Aus dem Institut für Genetik
Universität Hohenheim
Fachgebiet: Allgemeine Genetik

Prof. Dr. Gerald M. Rubin
Prof. Dr. Anette Preiss

Computational prediction of gene structure and regulation in the genome of *Drosophila melanogaster*

Dissertation
zur Erlangung des Grades eines Doktors
der Naturwissenschaften

der Fakultät II - Biologie
der Universität Hohenheim

von
Martin G. Reese
aus
Göttingen
2000

Die vorliegende Arbeit wurde am 31. März 2000 von der Fakultät II Biologie der Universität Hohenheim als “Dissertation zur Erlangung des Grades eines Doktors der Naturwissenschaften” angenommen.

Tag der mündlichen Prüfung:	4. Mai 2000
Dekan:	Prof. Dr. Anette Preiss
Berichterstatter, 1. Prüfer:	Prof. Dr. Anette Preiss
Mitberichterstatter, 2. Prüfer:	Prof. Dr. Gerald M. Rubin
Mitberichterstatter, 3. Prüfer:	Prof. Dr. Uwe Jensen

Again…

Felix qui potuit rerum cognoscere causas.

(VERGIL, GEORGICA **2**, 490)

This thesis is dedicated to the memory of my grandmother Maria Reese from Eichenberg.

Table of Contents

Tables

Acknowledgments

I thank my parents for their patience, support and constant encouragement, even from thousands of miles away. Thanks to Cristiana for understanding so much and being there when I needed her. Special thanks goes to my friend and closest co-worker David Kulp for his never-ending patience with me through so many difficult but rewarding projects. I would like to thank David Haussler, Michael Ashburner, Otto Ritter, Gary Stormo, Søren Brunak and Terry Speed and so many others for their scientific advice and encouragement for my research. Thanks to the GASP team, including George Hartzell, Uwe Ohler, and Nomi Harris, for a wonderful team experiment, which showed, how productive collaborations can be. Thanks also to our whole team at Neomorphic for their support throughout. I am very grateful to the entire staff of the BDGP for being such wonderful hosts to me. Particular thanks goes to my first advisor, Frank Eeckman, who made my work possible by inviting me to join his group in Berkeley. Finally, many thanks to Frank, Uwe, Nomi, David, Roger Hoskins, Andrew Martin, and Gert Riethmüller for very helpful editorial comments. Last but not least, I am indebted to my two thesis advisors, Gerry Rubin and Anette Preiss, who were even better advisors than a student could have hoped for. Their encouragement and support was tremendous.

Publications

Publications arising from this thesis

Reese, M.G., Kulp, D., Tammana, H. and Haussler, D. (2000). Genie - Gene finding in *Drosophila melanogaster*. *Genome Res*, **10**(4), 529-38.

Adams, M.D., *et al.,* (2000). The genome sequence of *Drosophila melanogaster*. *Science*, **287**(5461), 2185-95.

Reese, M.G. (2000). Application of a time-delay neural network to the annotation of the *Drosophila melanogaster* genome. Submitted *Comput Chem.*

Other publications

Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U. and Lewis, S.E. (2000). Genome Annotation Assessment in *Drosophila melanogaster*. *Genome Res*, **10**(4), 483-501.

Lewis, S.E., Ashburner, M. and Reese, M.G. (2000). Annotating eukaryote genomes. *Curr Opin Struct Biol.,* **10**(3), 349-54.

Reese, M.G., Eeckman, F.H., Kulp, D. and Haussler, D. (1997). Improved splice site detection in Genie. *J Comput Biol*, **4**(3), 311-23.

Kulp, D., Haussler, D., Reese, M.G. and Eeckman F.H. (1996). A generalized hidden Markov model for the recognition of human genes in DNA. *ISMB* **4**, 134-42.

Ashburner, *et al.* (1999). An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*. The *Adh* region. *Genetics* **153**(1), 179-219.

Zusammenfassung

Für die automatische Genomannotierung werden Methoden aus der Informatik benötigt, um die Masse dieser neu bestimmten Sequenzdaten verstehen und interpretieren zu können. Zu diesem Zweck wurden zwei neuartige Methoden entwickelt. Ihre Anwendung zur Analyse des Genoms von *Drosophila melanogaster* wird hier vorgestellt.

Die erste Methode ist ein Neuronales Netzwerk, das die strukturellen Eigenschaften und den Aufbau von eukaryontischen Promotorregionen modelliert. Dieses Modell wird dazu eingesetzt, Transkriptionsstartstellen in genomischen Sequenzen von *Drosophila* vorherzusagen. Das Modell ist ein *Time-Delay* Neuronales Netz, welches einen Spezialfall der *feed-forward* Topologie darstellt. Diese Netzwerkarchitektur wurde ursprünglich zur Erkennung gesprochener Sprache entwickelt und angewendet. Die *Time-Delay* Archtitektur erlaubt eine positionsunabhängige Verarbeitung der Eingabedaten und kann die variable Distanz zwischen den funktionellen Bindungsstellen eines Promotors modellieren. Diese variable Distanz ist ein wesentliches Charakeristikum des biologischen Transkriptions-prozesses. In einem ersten Schritt werden zwei voneinander unabhängige Neuronale Netze trainiert, eines für die Erkennung der TATA-Box und eines für die Erkennung der *Initiator*-Bindungsstelle. Dabei werden die statistischen Merkmale dieser beiden Muster anhand der vorhandenen Promoterdatensätze zunächst positionsspezifisch erlernt. Das *Time-Delay* Neuronale Netz integriert daraufhin die unabhängig trainierten Netzwerke und erfaßt so die potentiell nichtlinearen Abhängigkeiten und variablen Entfernungen zwischen den Sequenzen der beiden Bindungsstellen. Dieses Modell wurde anschließend auf einer Testmenge aus sogenannten Kernpromotoren (*core promote*rs) angewendet, welche die unmittelbare Region um den Transkriptionsstart umfassen (+40..-11). Dabei ergab sich, dass die Klassifikationsrate bereits existierender Verfahren (statistische Klassifikatoren bzw. einfache Neuronale Netze) verbessert werden konnte.

Das *Time-Delay* Neuronale Netz wurde in einem Computerprogamm, NNPP, implementiert. Die Besonderheit des Programms besteht darin, daß der Transkriptionsstart positionsgenau vorhergesagt werden kann. Außerdem wird darauf eingegangen, daß das Programm nur eine kurze Berechnungszeit benötigt, und wie es leicht in einem kompletten Annotierungssystem verwendet werden kann. Als Beispiel wird NNPP in das im folgenden beschriebene GENIE-System integriert und die resultierenden Ergebnisse präsentiert. Die Vorhersagegenauigkeit des Programs ist besser als andere, sogenannte signalbasierte Methoden, und mindestens ebensogut wie sogenannte inhaltsbasierte Methoden.

Ein Test auf einer großen Menge genomischer Sequenzen ergab eine Vorhersage-genauigkeit von 75 Prozent (69 von 92 Promotoren wurden erkannt), mit einer falsch positiven Vorhersagerate von 1/547 Basen. Für das in GENIE integrierte System fällt die Vorhersagerate auf 32,6 Prozent (30/92), aber dafür ist die falsch positive Rate erheblich verbessert (1/16,729 Basen). Die positive Vorhersagerate ist niedriger, da durch die Restriktionen im GENIE-System die Vorhersage auf die Region vor einem vorhergesagten Gen eingeschränkt ist.

Das zweite vorgestellte System ist ein Wahrscheinlichkeitsmodell der eukaryontischen Genstruktur und deren spezifischer Eigenschaften für *Drosophila melanogaster*, ein generalisiertes Hidden-Markov-Modell (GHMM). Sowohl für die einzelnen Signale wie den Transkriptions- und Translationsstart und die Spleißstellen, als auch für Regionen wie Exons, Introns und die intergenische Region, werden jeweils Wahrscheinlichkeiten berechnet. Modellparameter für diese Regionen und Signale werden extern bestimmt. Für codierende Bereiche werden beispielsweise die Verwendung und Bevorzugung bestimmter Codons als charakteristische Merkmale eingesetzt. Als Sensoren für Signale sind sogenannte Gewichtsmatrizen integriert, welche die Auftrittswahrscheinlichkeiten der Nukleotide an den einzelnen Positionen beschreiben. Die Teilmodelle werden unabhängig mit einem Satz von repräsentativen, bekannten *Drosophila*-Genen trainiert. Zusätzlich zu diesen Modellen, welche die statistischen Eigenschaften von Genen ausschließlich anhand von bekannten Gensequenzen erlernen und daher "*ab initio*" genannt werden, wird eine neue Methode für die Integration von EST-Sequenzen vorgestellt. Diese Methode zwingt das GHMM dazu, ein Intron vorherzusagen, wenn die beiden benachbarten Sequenz-bereiche von ein und derselben EST-Sequenz überlappt werden. Dafür wird die gesamte EST-

Datenbank für *Drosophila* benutzt. Sollte eine 5' und eine 3' EST-Sequenz von dem gleichen cDNA-Klon vorhanden sein, so wird diese Information benutzt, um den Start und das Ende eines Gens zu markieren.

Die Entwicklung des GENIE-Computerprogramms wird erläutert. Dieses Program kann sowohl mehrere komplette wie auch zum Teil unvollständige Genstrukturen innerhalb derselben Sequenz von *Drosophila* identifizieren. Weitere besondere Neuheiten des Systems, neben der Integration der EST-Information, sind die Integration von Ähnlichkeiten zu Proteinsequenzen in anderen Organismen und die Bewertungsmethode an sich, die eine Gesamtwahrscheinlichkeit anhand aller Teilmodelle für ein bestimmtes Gen berechnet, einen sogenannten *gene parse*. Außerdem ist das Modell besonders flexibel, so daß verschiedene externe Untermodelle sehr leicht integriert werden können. Dieses wird am Beispiel des NNPP-Programmes gezeigt.

Um die Genauigkeit eines *ab-initio*-Genvorhersageprogramms wie z. B. GENIE zu messen, und um die Nützlichkeit eines solchen Systems zu demonstrieren, wurde ein internationales Experiment durchgeführt. Für dieses **G**enome **A**nnotation **A**ssessment **P**roject (GASP) wurde eine besonders detailliert untersuchte Genomregion ausgewählt -- die *Adh*-Region von *Drosophila*, welche 222 annotierte Gene enthält. Die eingesandten Computervorhersagen von den verschiedenen Gruppen wurden anhand von hochqualitativen, unveröffentlichten cDNA-Sequenzen, von denen man die genaue Genstruktur in der DNA leicht direkt ableiten kann, ausgewertet. Eine zweite, vollständigere Menge von Genen für die Auswertung wurde aus einer kürzlich durchgeführten, detaillierten Studie dieser Region übernommen. Im Rahmen des GASP-Experimentes hat die *Drosophila*-Version des GENIE-Programms, welches EST-Sequenzen als zusätzliche Information für die Erkennung von Genstrukturen benutzt, besonders gut abgeschnitten. Über 95 Prozent der codierenden Nukleotide in der Region konnten korrekt bestimmt werden. Außerdem überlappen 90 Prozent der gesamten, vorhergesagten Genstrukturen zumindest teilweise die vorher annotierten 222 Genen aus der Studie. Es wurden 26 zusätzliche Gene gefunden, von denen einige wahrscheinlich falsch positive Vorhersagen sind. Die von GENIE ermittelten Enden der Exons, die Spleißstellen, wurden als besonders zuverlässig eingeschätzt: 77 Prozent der durch cDNA-Sequenzen bestätigten Exons wurden korrekt von GENIE vorhergesagt. Von den 43 annotierten Genen aus der ersten Testmenge wurden 19 in der

Gesamtstruktur absolut exakt gefunden. Dieses bedeutet, daß für etwa 50 Prozent der Vorhersagen die Proteinsequenzen komplett und korrekt bestimmt werden können.

Eine weitere, bedeutende Anwendung des NNPP-Programmes zur Vorhersage von Transkriptionstartstellen wird gegen Ende der Arbeit vorgestellt und diskutiert. In dieser Anwendung wurde ein potenzieller bisher unbekannter Transkriptionsstart und damit außerdem eine verlängerte kodierende Gensequenz für das *C. elegans* Gen *unc-86 in silico* von NNPP vorhergesagt. Die Vorhersage konnte mittlerweile in verschiedenen biologischen Experimenten, darunter die neueste Primer-Extension-Methode RACE sowie verschiedene cDNA Library Screening Methoden mit ausgewählten Primers aus der Transkriptionsregion, bestätigt werden.

Die generellen Anwendungsmöglichkeiten und Anpassungmöglichkeiten der entwickelten Methoden werden am Ende erläutert. Wie die Ergebnisse zeigen, können die hier entwickelten, analytischen Methoden auch für andere eukaryontische Genome, wie z.B. das menschliche Genom, angewendet werden.

Die vorgelegte Arbeit ist somit eine umfassende Studie über die Entwicklung und Anwendung neuer Technologien zur automatischen Erkennung von Genstrukturen und der Regulation von Genen im Genom von *Drosophila melanogaster*.

Thesis abstract

Computational methods for automated genome annotation are critical to understanding and interpreting the bewildering mass of genomic sequence data presently being generated and released. Two such methods, both novel, have been developed and their application for analysis of the *Drosophila melanogaster* genome are presented. The first represents a neural network model of the structural and compositional properties of a eukaryotic core promoter region. It is applied to the problem of predicting transcription start sites in genomic sequences of *Drosophila*. The model uses a time-delay architecture a special case of a feed-forward neural network that originally was developed for speech recognition. The structure of this model allows for variable spacing between functional binding sites, which is known to play a key role in the transcription initiation process. Individual neural networks for the recognition of the TATA box and the initiator binding sites were trained specifically to recognize the peculiar position of these consensus sequences in a series of eukaryotic promoters. During this training the statistical properties of the nucleotides in the binding sites were learned. The combined time-delay neural network model then incorporates these individual networks and captures potentially important dependencies between the individual binding sites. Application of this model to a test set of core promoters not only gave better discrimination of potential promoter sites than previous statistical or neural network models, but also revealed indirectly subtle properties of the transcription initiation signals which allowed to deduce certain aspects of their biochemical function. The development of a computer program (NNPP), that identifies potential transcription start sites in genomic DNA using the above mentioned time-delay neural network model is described. The uniqueness of the program consists in the ability to recognize precisely the position of a transcription start site for a given gene. Its fast running time and its capacity to be ported into Genie, shows that it can easily be integrated in a whole genome annotation system. The accuracy of the program is substantially better than

similar methods and at least as good as content-based methods that require information from larger genomic regions.

When tested in the *Adh* region of the *Drosophila* genome, the stand-alone NNPP program gives a recognition rate of 75 percent (69/92) with a false positive rate of 1/547 bases. When integrated into Genie, the recognition rate drops to 32.6 percent (30/92) with a much-improved false positive rate of 1/16,729 bases. The recognition for the integrated system is lower because of the constrained prediction of a transcription start site upstream of a potential gene.

The second system introduced is a probabilistic model of the gene structural and compositional properties of *Drosophila* genomic DNA. The novel statistical model is a generalized hidden Markov model whose architecture incorporates probabilistic descriptions of signal sensors for start of transcription and translation and splice sites, as well as content sensors for exons, introns and intergenic regions. Model parameters for these content and signal sensors models are derived for codon preference and codon usage as well as the nucleotide compositions for the consensus sites for splice sites and start codons. These submodels are trained on a collected representative set of known *Drosophila* genes. In addition to these *ab initio* gene finding statistics a novel method to integrate EST sequence alignments through constraints is described to predict introns flanked by coding exons. Specifically for *Drosophila* the existing collection of 5' and 3' EST sequences is used as a constrain for the gene structure model to cluster predicted exons into one contiguous gene between an alignment of 5' and 3' EST sequences from the same cDNA clone.

Development of the Genie computer program is described, which identifies complete gene structures in *Drosophila*. New features of the program include the use of alignments of EST sequences, the integration of homology information derived from protein sequences of other organisms, and the integrated probabilistic scoring of a gene parse. In addition this program allows for the easy integration of specific submodels such as the NNPP promoter model and predicts consistently genes on both DNA strands. To assess the accuracy of an *ab initio* gene finding program such as Genie and other similar systems and their usefulness for annotating the genome of *Drosophila melanogaster* an international experiment was initiated. For the experiment - known under the acronym GASP (Genome Annotation Assessment Project), a large, well-

characterized sequence contig was chosen: the *Adh* region in *Drosophila*, which has 222 annotated genes. Computational predictions, made by the participating groups, were evaluated using two standards, one based on previously unreleased high quality full-length cDNA sequences and the other derived from a set of annotations generated as part of an in-depth study of the region by a group of *Drosophila* experts. The *Drosophila* version of Genie incorporating EST alignments, GenieEST, was one of the best programs when tested on this standardized set of genes. Over 95 percent of the coding nucleotides in the region were correctly identified and Genie's gene assignments overlapped with 90% of the previously 222 annotated genes. In addition, twenty-six novel genes were predicted some of which might be false positives. Exon boundary assignments made by GenieEST were judged to be substantially better than other evaluated methods. 77 percent the cDNA confirmed exons were correctly predicted by Genie. In the gene assembly class, GenieEST correctly predicted nineteen out of the 43 annotated genes. Thus one can reasonably expect that of almost 50 percent of the genes predicted by Genie the complete and correct protein sequence can be derived.

Another typical application of the NNPP promoter prediction program is discussed where a novel potential transcription start site and an extended coding region for the *unc-86* gene in *C. elegans* was predicted by NNPP. The predictions were subsequently verified by various experiments including the primer extension method RACE and cDNA library screening using selected primers.

In general, applicability and portability of the developed methods to various other organisms are discussed. As the data shows, the developed analytical methods are well suited to use in other eukaryotic genomes, including human.

The presented works can be regarded as one of the first intensive studies that applies novel gene finding and regulation technologies for the identification of complex genes structures and regulation in the genome of *Drosophila melanogaster*.

Chapter 1 Introduction

Recent advances in sequencing technology are making the generation of whole genome sequences commonplace. Capillary sequencers speed the production of raw data. Changing tactics from traditional mapping and sequencing clones in series to an integrated simultaneous mapping and sequencing approach (whole genome shotgun) has significantly reduced the amount of time it takes to completely sequence a genome. These improvements in genomic sequencing are possible because of software advances that fully exploit mapped clone constraint data and directly attack the problems that repetitive sequences cause during sequence assembly.

At present several very large-scale genomic sequencing projects are complete or are expected to complete within a few months. These initial genome sequences are from key model organisms in genetics and include five eukaryotes, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Arabidopsis thaliana*, as well as draft human sequence. In a few years sequencing new genomes and individuals will become routine practice. This raw data is not immediately useful and interpreting it places major demands on the field of computational biology.

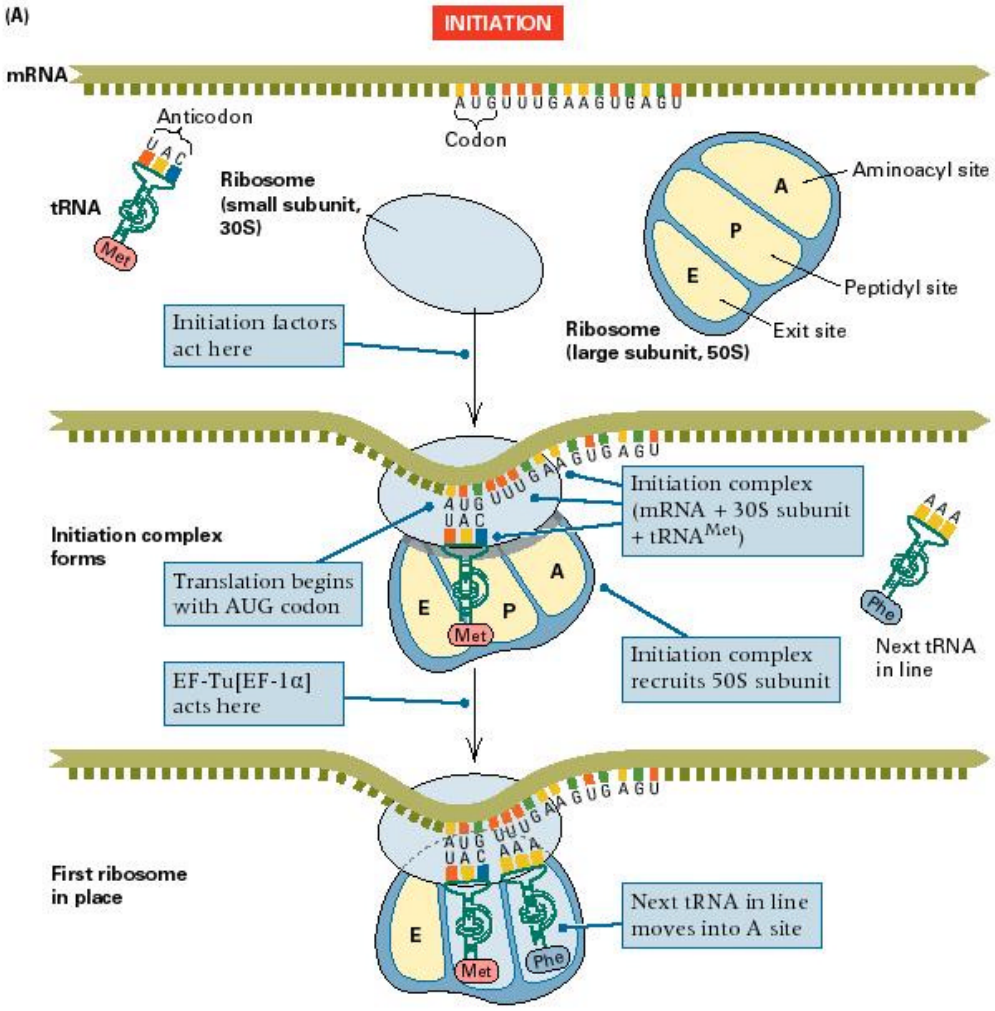
The discipline of computational biology can be described as the intersection of genetics, molecular biology, structural biology, molecular evolution, physics, mathematics, computer science and engineering. Its goal is often to use computational, statistical or mathematical methods to understand the relationships between sequence, structure, evolution, biological function, molecular behavior and genetics. Traditionally the focus of the field has been in the prediction of protein structures from the raw protein sequence, in evolutionary studies of complete genomes or individual proteins or protein families, in correlation studies of the global structure of genomes, in mapping and sequencing support technologies and statistics; more recently applications have

arisen in the study of mRNA expression levels and in the area of molecular genetics and epidemiology.

This thesis work addresses a significant open problem in computational biology as well as in genomics in general - genomics defined as the science of studying complete genomes -, the problem of computational genome annotation. Genome annotation is a rapidly evolving field in genomics made possible by the large-scale generation of genomic sequences and driven predominantly by computational tools. The goal of the annotation process is to assign as much information as possible to the raw sequence of complete genomes with an emphasis on the location and structure of the genes. This can be accomplished by *ab initio* gene finding, that is through the application of statistical modeling of genomic sequence alone, or in consort with *homology-based* gene finding, in which genes from other organisms are aligned to the genomic sequence.

This work specifically addresses two important problems, namely the identification of the precise exon-intron structures of genes and the recognition of regulatory promoters in higher eukaryotic (especially *Drosophila melanogaster*) genomic DNA sequence. The gene identification problem is a bottleneck in any genome annotation process because it is the foundation of any other subsequent characterization of genes. Localization of the beginning of transcription plays a major role in the immediate next step of annotation regarding the regulatory mechanism of a gene. Besides the practical importance of automated computer methods for annotation there is also an intrinsic biological interest in both problems. To find good solutions for both problems we are challenged to precisely study the statistical properties in the sequence, which are dependent on the fundamental biochemical processes of transcription, translation and RNA splicing. (For a schematic overview of the fundamental biological processes involved in transcription, splicing and translation see Figure 1-1) Our modeling allows us to hypothesize about the mechanical properties of these underlying processes.

Figure 1-1: From transcription initiation to translation (from Hartl and Jones, 1998)



The approach taken in my thesis is to develop a sophisticated neural network using a specialized network architecture based on the individual protein binding sites in a core promoter. For the gene identification problem in *Drosophila*, a probabilistic model of gene structure expressed as a generalized hidden Markov model is developed. Both models are then applied to the problems of transcription start site and gene structure identification in *Drosophila melanogaster* in two computer programs called NNPP (**N**eural **N**etwork for **P**romoter **P**rediction) and Genie.

This work is an interdisciplinary approach typical of computational biology where the subject of the study is biological, results of biological interest are obtained and techniques are applied from many other fields such as stochastic models from statistics,

neural networks from speech recognition, dynamic programming algorithms from computer science and computer engineering and integration systems from information sciences.

The playing field and science of this analysis is the genome of *Drosophila melanogaster,* which has been studied genetically over the last 100 years and is an excellent model organism for human. Traditionally, small-scale studies of isolated genes carried out in an individual researcher's laboratory use a combination of computational and experimental methods that permit very detailed descriptions of genes and its features. They offer a narrow but deep view. In contrast, the best current results from the annotation of large eukaryotic genomes such as *Drosophila* provide a complete perspective and overview on the entire genome, albeit superficially. They offer a broad but shallow view. At present the annotation of large-scale sequences is a compromise, but ideally the aim is to have both breadth and depth in our description of the genome. Computational tools are strongly needed to improve the information derived from these genomes.

Almost 100 years of scientific research has passed since W.E. Castle and his colleagues introduced *Drosophila melanogaster* as a model organism for biological studies (Kohler, 1994). From the very beginning the fruitfly has dominated research in genetics. In November 1999 *Drosophila* joined the elite group of completely sequenced organisms (Rubin *et al*., 2000). There are many reasons why a complete genome is so important for future research. From a practical point, it will be of great benefit to all scientists around the world that study particular genes. From a theoretical point, the finished genome will give a complete description of all the proteins in *Drosophila* - assuming all the genes are identified - and therefore be a blueprint for the entire organism. Studies on intergenic regions, number of genes, number of exons, structural organization and transposon distributions will be possible. Finally the detection of entirely novel protein families in invertebrates, or specific to *Drosophila melanogaster,* will be detected.

To prepare tools for this event and to study the open problems in the analysis and interpretation of long genomic sequences, early on in the *Drosophila* genome project the 2.9Mb *Adh* genomic region was selected as a test bed (after the *Adh* gene, which codes for for the *alcohol dehydrogenase* protein). This region was already well

characterized by conventional genetic analyses. This chromosome region is defined as the 69 polytene chromosome bands from 34C4 to 36A2 on the chromosome arm *2L*, which is the region between (and including) the previously known genes *kuzbanian* (*kuz*) and *dachshund* (*dac*). Genetic studies of the chromosome region began with the recovery of the Adh^{-} deletion (Grell *et al*., 1968).

In a very interdisciplinary effort involving many different laboratories and scientists with different skills this region was sequenced and was manually annotated over several years. In fall 1999, Ashburner *et al.* (Ashburner *et al*., 1999) published a definitive annotated sequence. Nearly 3 megabases (Mb) were sequenced from a series of overlapping P1 and BAC clones as a part of the Berkeley *Drosophila* Genome Project (BDGP) (Rubin *et al.*, 1999) and the European *Drosophila* Genome Project (EDGP) (Ashburner *et al*., 1999). This sequence is believed to be of very high quality with an estimated error rate of less than 1 in 10,000 bases, based on PHRAP quality scores (Ewing & Green, 1998). Computational analysis in conjunction with expert knowledge of the sequence predicted 218 protein-coding genes, 11 tRNAs, and 17 transposable element sequences. The gene density of protein-coding genes is one per 13 kilobases (kb). A detailed analysis of this region can be accessed through the BDGP web site (http://www.fruitfly.org/publications/*Adh*.html) as well as in Ashburner *et al.* (1999). This region is thought to be typical and therefore a good test bed to draw conclusions for the entire genome. This region provided the substrate for annotation in this thesis.

Besides this focused effort within the BDGP and EDGP, a more global effort has been initiated to assess and evaluate computational annotation methods in a global experiment performed on this *Adh* region before the annotations were published. A short summary in connection with the evaluation of my work is given in the thesis Chapter 4.

After the sequence is finished we are faced with the problem of feature identification. The types of features that can be detected and described in the genomic sequence are the location of the protein coding genes; the structure of those genes (including untranslated regions and control elements in addition to the exon-intron structure for all possible transcripts); the probable translations of every transcript into a protein product; the location of the repetitive sequences and their nature, and the

location of the genes encoding non-coding RNAs. This is only a partial list and can easily be expanded. The identification of these essential elements of the genomic sequence is a necessary basis for annotation.

There are two major classes of technique for the prediction of genes - *ab initio* methods and homology-based methods. In prokaryotes, and in some simple eukaryotes such as *Saccharomyces cerevisiae*, genes normally have single continuous open reading frames and short intergenic regions separate adjacent genes. By contrast genes in most eukaryotes may be very complex, with many exons and with introns that may be ten's of kilobases in length. Eukaryotes also tend to have more complex non-coding 5' and 3' exons and alternatively spliced products. In addition, complex relationships between genes may be quite frequent, e.g. genes that are contained within the introns of other genes and adjacent series of very related genes. The consequence is that any *ab initio* method must combine the prediction of gene components - exons, introns, splice sites etc. - with a model of how these components may be assembled into a gene.

This work here presents a combination approach that integrates *ab initio* statistics and second hand information from mRNA and homologous sequences. The promoter localization approach is mostly *ab initio* but the integrated model into the gene finding system puts the *ab initio* model in context with cDNA alignments in the gene finding system.

Identification as described in this work leads to the third major problem - the characterization of genes. This characterization must be done in several ways: in terms of the relationships between the sequences of the elements and other sequences (both within the genome being annotated and with other genomes), in terms of the structure of the elements (e.g. the protein domains of predicted proteins), and in terms of the predicted function of the elements (e.g. what inferences can be drawn concerning the biological function of a predicted protein). Functional characterization of genes has not been the focus of this work and therefore I refer to extensive literature in this field.

Chapter 2 Promoter prediction using time-delay neural networks (NNPP)

In this chapter, an introduction to the transcriprediction using time-delay neural networption initiation process and a short historical review of existing methods for modeling this biological process is given (Section **2.1**). This is followed by a description of the existing promoter data sets (Section **2.2**) and their computer representation (Section **2.3**) during neural network training. Section **2.4** introduces neural networks which is followed in Section **2.5** by a detailed description of the time-delay architecture and algorithm used in this work. Section **2.6** summarizes the implementation of the final computer program NNPP.

2.1 Background

One of the challenges in the field of computational biology and especially in the area of computational DNA sequence analysis is the automatic detection of promoter sites. Promoter sites typically have a complex structure consisting of multiple functional binding sites for proteins involved in the transcription initiation process. Therefore, I have focused in this work on detection of the core-promoter region, a sequence region spanning up to 50-60 basepairs upstream and 10-15 basepairs downstream of the transcription start site (TSS). This is a subset of the promoter region which spans 300 basepairs upstream and 50 downstream of the TSS. The primary goal was to build an automatic computer program to localize the TSS in a genomic DNA sequence as precisely as possible. This program has three equally important applications:

1. In genome-wide annotation using gene finding programs such as Genie, the automated partitioning of exons among several genes is very difficult.

Successful promoter recognition promises to correctly identify the beginnings of genes, thereby enabling a major advance in multi-gene recognition;

2. Knowledge of the TSS facilitates locating the correct initiation codon, resulting in better coding gene prediction;

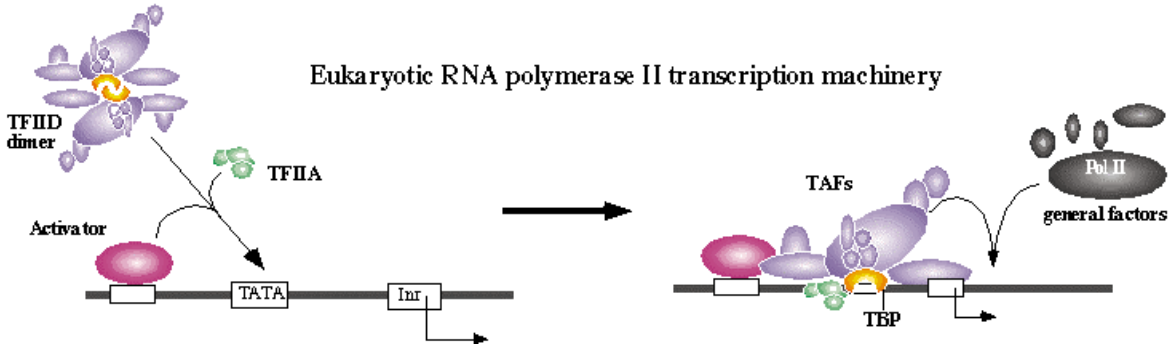
3. Applying automated TSS localization in conjunction with cDNA alignments to genomic DNA can categorize a cDNA sequence as "full-length" if its 5' sequence end localizes close to a predicted TSS. This is very important, because biologically confirming full-length cDNAs is a very difficult and labor intensive task;

Better TSS recognition will also lead to a better understanding of the structure and mechanisms of regulatory elements and the entire gene regulation process. Precise TSS annotation will narrow the search space for regulatory elements such as *cis*-regulated binding sites. With the recent flood of gene expression data, especially from DNA microarray experiments, understanding regulation of transcription and identification of TSS's will become more and more important.

2.1.1 Eukaryotic transcription initiation

Transcription is initiated by specific interactions between several transcription factors, RNA polymerase II, and the DNA sequence in the promoter region. These biochemical processes are currently the focus of intense investigation. Recent progress is reviewed in (Burley & Roeder, 1996; Kornberg, 1999; Pugh, 1996; Pugh & Tjian, 1992; Roeder, 1996; Yokomori *et al*., 1998).

Figure 2-1: Eukaryotic RNA polymerase II transcription machinery (from the laboratory of B.F. Pugh, 2000)



Experimental and theoretical studies of eukaryotic promoter regions - most of them performed in *Drosophila melanogaster* - have shown that promoters have a complex sequence structure reflecting the complicated transcription initiation process (see Figure 2-1 for a schematic view). A so-called pre-initiation complex (PIC) recognizes the core promoter and initiates transcription. The PIC includes the general transcription factors (GTFs) TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH and RNA Polymerase II. Each of these factors may comprise multiple protein complexes. TFIID, the most well-known and well-studied complex, consists of the TATA-binding protein (TBP) and several TBP-associated factors (TAFs). Very recently, the three dimensional structure of the binding complex around the human multisubunit TFIID has been obtained by electron microscopy and image analysis (Andel *et al*., 1999). This work showed that TFIID forms a TFIID-IIA-IIB complex to assemble the initiation complex at a eukaryotic core promoter. TBP, the DNA binding protein in this complex, binds site-specifically to the so-called TATA box. This TATA box, reviewed in (Breathnach & Chambon, 1981; Bucher, 1990; Conaway & Conaway, 1993; Penotti, 1990; Pugh & Tjian, 1992)) is the most conserved sequence motif in the core promoter region. It is located 15-25 bases upstream of the TSS in metazoans. Spanning the TSS, including the cap site, is the so-called “initiator” (Inr) (O'Shea-Greenfield & Smale, 1992; Smale & Baltimore, 1989). It is a much less well-conserved sequence and is therefore a much weaker signal than the TATA box.

Other statistically significant motifs in the eukaryotic core promoter that are not present in vertebrates are the GC box (Bucher, 1990; Lisowsky *et al*., 1999) and the CAAT box (Bucher, 1990; Bucher & Trifonov, 1988). These sites occur mostly outside the core region, but the rules governing the exact location are not well understood. In all conserved core promoters the relative positions of the elements and the TSS are highly variable and some elements may be entirely absent (for a review on TATA-less promoters see (Smale, 1997; Wiley *et al*., 1992)). For further details on the biological process see Fickett and Hatzigeorgiou (1997) and references therein.

2.1.2 Computational promoter prediction in prokaryotes

The first *in silico* promoter studies concentrated on prokaryotic promoters, which have less complex structures than their eukaryotic counterparts. Hawley and McClure (1983) pioneered systematic computational analysis of *E.coli* promoters. They studied

168 promoter regions and established a consensus sequence for prokaryotic promoters. Harley and Reynolds (1987) extended this work by using Hawley and McClure's early work to identify 263 additional *E.coli* promoters. Nakata *et al.* (1988) then used this compilation of promoters and applied the "perceptron" algorithm (Minsky & Papert, 1969) to build a minimal neural network consisting of only one input layer and one output layer. This work can be seen as the first application of discriminative training to the problem of promoter recognition. Demeler and Zhou (1991) extended Nakata's work by formally introducing a multi-layer neural network architecture to the problem of DNA motif detection and showed its strong classification power using *E.coli* promoters. Demeler and Zhou's neural networks were trained using the backpropagation algorithm. O'Neill (1991; 1992) extended the neural network approach by pioneering the application of trained neural networks to scanning long, contiguous genomic sequence. O'Neill was also the first to explicitly deal with the variation in spacing of the promoter binding sites (-35 and -10 box in prokaryotes) by modeling and training different neural networks for different spacing classes. All these early prediction methods were then summarized and assessed by Horton and Kanehisa (1992), who obtained a stronger classification with neural networks than with conventional statistical methods. These early studies in prokaryotes suggested that both sequence specificity at the binding sites and distances between sites play a key role in the initiation process. A more recent computational study of prokaryotic promoters by Pedersen *et al.* (1995) showed the ability of computational models to identify novel sequence motifs and to find new signals that are regularly spaced along the promoter region. Pedersen *et al.* show evidence that the spacing of weaker signals corresponds to the helical periodicity of DNA.

2.1.3 Computational promoter prediction in eukaryotes

One of the first statistical studies of RNA Polymerase II promoter regions in eukaryotes was performed by Bucher (1990; 1986), who analyzed functional promoter sites from different eukaryotes and built statistical weight matrices for each individual element, such as the TATA box, Inr site, CAAT box and the GC box. The weight matrices were based on counts of a specific nucleotide at a fixed position. Penotti (1990) has extended this work, using human promoter sequences from the EMBL primate database. He showed statistically significant differences in the human derived TATA box weight matrices of Bucher's vertebrate- and virus-derived collection.

Penotti's "information content measure" showed significant sequence signals for the Inr site for the human sequences.

Matis *et al.* (1995) computed sequence weight matrices for a large collection of unrelated TATA box containing promoters. In addition to the sequence profiles, they calculated and analyzed distance distributions between various functional elements. They combined the results of the statistical studies into a "backpropagation feed-forward" neural network system (defined below) (Rumelhart *et al*., 1986b), which assigns scores to potential promoter regions. The results from the neural network were then integrated into the gene finding system GRAIL2 (Uberbacher & Mural, 1991) to reduce the number of false positives.

Prestridge (1995) developed a computer program, PROMOTER SCAN, which utilizes promoter recognition profiles derived from a transcription factor database such as TRANSFAC (Wingender *et al*., 2000; Wingender *et al*., 1996). For the final prediction, the promoter recognition profile is combined with the Bucher TATA box weight matrix (1990). This has been the state-of-the-art program.

The "general data study" by Larsen *et al.* (1995) revealed an additional CT- signal positioned on average seven nucleotides downstream from the TSS for which no binding factor is yet known. This study used neural networks to derive new signals.

In 1997, Fickett and Hatzigeorgiou (1997) published an excellent overview of the status of eukaryotic promoter recognition algorithms. Besides giving a great introduction and overview of the biological process, they compared various programs, including an early version of the following TDNN algorithm, on a standardized data set. They found recognition rates on the order of 13%-54% of known promoters, and false positives on the order of one per kilobase (NNPP 's results in this review are discussed below). Fickett and Hatzigeorgiou concluded that the problem of eukaryotic promoter prediction is complex and far from solved.

Following the Fickett review, interest in this field shifted from general TSS prediction to tissue specific promoter prediction (Frech *et al*., 1998; Klingenhoff *et al*., 1999; Wasserman & Fickett, 1998). These models have much lower false positive recognition rates but they do not address the problem of genomic promoter recognition.

Recent work on general promoter prediction was presented by Ohler *et al.* (Ohler *et al*., 1999) who developed a novel content-based approach based on interpolated Markov chains, which they have later extended to stochastic segment models (Ohler *et al*., 2000). This method distinguishes promoter sites from coding and non-coding regions. The approach is very promising because it also allows the detailed study of significant binding site patterns in the specific submodels of the system. Ohler *et al.* trained their program on *Drosophila* data as well, and results were reported in the paper by Reese *et al.* (2000).

All methods developed to date have been plagued by inconsistent performance and inability to predict the exact position of TSS's. Given the nature of genomic sequence in higher organisms such as *Drosophila* and human, with large introns and intergenic regions a very specific algorithm with the ability to precisely predict the position of the TSS is needed.

2.2 Core promoter data sets

We selected two representative data sets, one from the Eukaryotic Promoter Database, EPD (Cavin Périer *et al*., 2000), hereafter called the "promoter set", and one from Genbank, hereafter called the "gene set" (both datasets can be found in (Reese & Ohler, 1999)).

The promoter set consists of 429 unrelated vertebrate promoter regions (-350 to +50) with a sequence identity between any two sequences of less than 25% (Appendix B). The set consists of 37.4% human, 23.4% mouse and 12.1% chicken (*gallus*) sequences. The remaining 27.1% are from various other eukaryotes including *Drosophila* and eukaryotic viruses.

The gene set is a collection of non-redundant human genes representing typical coding regions in metazoans such as *Drosophila* and human. This data set was also used in the training of Genie (see Chapter 4 and Appendix C).

The combined data were split into a training set and a test set in the following way. The training set consists of 3,300 sequences containing 51 bases each. Of these, 300 sequences contain randomly selected promoter regions from the promoter set, while the other 3,000 contain sections of coding sequence from the gene set. The 300 promoter

sequences contain the regions from -40 to +11, where +1 is the TSS as annotated in EPD. The test set consists of 1,129 sequences of 51 bases. 129 of these include eukaryotic promoters and 1,000 contain regions of random coding sequence. All reported results are 4-fold cross-validated.

We report as true positives those results, where given a cutoff threshold, the network predicts a promoter in a sequence that contains a promoter. False positives are those results where given the same threshold, the network predicts a promoter in a sequence representing random coding sequence. For example, a false positive rate of 3.8% for the test set would indicate that for a given threshold on the output unit, the network predicts 38 promoters in the 1,000 sequences containing no promoters.

2.3 Input data representation

It has been shown in earlier applications of neural networks to nucleic acid sequence analysis (Brunak *et al*., 1991; Demeler & Zhou, 1991; Farber *et al*., 1992; O'Neill, 1992; Qian & Sejnowski, 1988) that the orthonormal coarse code of fourth dimension (A: 1 0 0 0; C: 0 1 0 0; G: 0 0 1 0; T: 0 0 0 1) gives the best results. This data representation uses a unitary coding matrix with identical *Hamming distance* between each pair of vectors. This guarantees that there is no correlation between the coding vectors of different nucleotides that could bias the learning procedure. True promoters in the training and test set are coded as “1”, and negative examples, which means non-promoters, are coded as “0” in the output of the neural network.

2.4 Neural networks

Neural networks are machine-learning techniques that were developed mainly in the field of signal and speech recognition. The early “neural networks” were inspired by concepts from neuroscience. The first to introduce the notion of a simple model of a neuron as a binary threshold unit were McCulloch and Pitts (1943). Specifically, the model neuron computes a weighted sum of its inputs from other units, and outputs a one or a zero according to whether this sum is above or below a certain threshold. In the early sixties the group of Frank Rosenblatt introduced the first form of a connected network, the so-called perceptron. Rosenblatt (1962) proved the convergence of a learning algorithm for the simplest class of perceptron, a two layer neural network. Minsky and Papert pointed out the severe limitations of the perceptron in their book

Perceptron (Minsky & Papert, 1969). In particular, they provided a proof that a single perceptron could only solve linearly separable problems, a very minor and uninteresting class of problems. The XOR is the best-known example for such a non-linearly separable problem. Interest in neural networks immediately declined dramatically.

Werbos (1974) made a very significant contribution in the early seventies. He described the backpropagation algorithm, a method to adjust the weights connecting units in successive layers of multi-layer perceptrons. The significance of backpropagation was overlooked for a whole decade until Rumelhart, Hinton and Williams (1986a; 1986b) independently rediscovered it. The name perceptron is no longer used and multi-layer networks are now called neural networks.

Hertz *et al.* (1991) give an excellent introduction and overview of the theory of neural computation. For an overview of applications in computational biology, for example, see my earlier work (Reese, 1994).

2.5 Time-delay neural networks

For promoter modeling, a special neural network is chosen, the time-delay neural network (TDNN) architecture developed by Waibel *et al.* (1989). This architecture was originally designed for processing speech sequence patterns in time series with local time shifts. The usual way of transforming sequence patterns into input activity patterns is the extraction of a subsequence using a fixed window. This window is shifted over all positions of the sequence and the subsequences are translated into input activities. The network produces an output activity or score for each input subsequence.

The following two promoter specific features have to be learned:

- The network has to recognize subsequences that may occur at non-fixed positions in the input window. Therefore the network has to learn that the subsequence is a feature independent of shifts in its position.

- The network has to recognize features even when those features appear at different relative positions. This situation arises in cases where different subsequences occur in the input window with different relative distances. This happens very frequently in genomic sequences when one or more elements (nucleotides) are inserted or deleted in a given promoter.

The TDNN architecture addresses these problems by imposing certain restrictions on the network topology and by the way in which weights are updated. Hidden units are connected to a limited number of input units that represent a consecutive pattern in the input window. These hidden units have a *receptive* field, that is, they are only sensitive to a part of the input window. The important restriction is that the same *receptive field* has to be present at each position in the input exactly once. If the input window contains, for example, ten positions and a *receptive field* covers a subsequence of three positions, there must be eight hidden units with the same *receptive field*. Since the corresponding weights in all copies of a *receptive field* are forced to have the same values, these hidden units are said to have *linked receptive fields*. In neural network terminology this is also known as *weight sharing*. Each hidden unit is called a *feature unit* because it will recognize a certain feature in the input window irrespective of its relative position. During learning, the partial derivatives of corresponding weights in *linked receptive fields* are calculated separately since these hidden units with their *receptive fields* at different positions in the input window get different activation. To adapt a *receptive field*, the weight update is averaged over all copies of a weight. This average update is then applied to all copies of that weight. In this way, it is ensured that the copies of a *receptive field* remain identical for a given feature. In the basic TDNN architecture the hidden layers (feature units) are connected to the output layer in a standard feed-forward way. Training is performed using a modified backpropagation algorithm.

There are several successful applications of TDNNs in speech recognition (Waibel et al., 1989) and the recognition of handwritten characters (Lang & Waibel, 1990). These references include a detailed description of the time-delay architecture.

2.6 Implementation of the core-promoter time-delay neural network model (NNPP)

Using the time-delay architecture described in section 2.5, two distinct neural networks, one for the TATA box and one for the Inr, were trained. We selected an input window of 30 bp (-40 to -10) for the TATA box neural network and a window of 25 bp (-14 to +11) for the Inr network. The window sizes were selected so that the consensus sequences for both binding sites are included. The two signals occur at varying distances relative to the TSS, which is used as the alignment point for the promoter

sequences in the EPD database. This implies that we do not wish to utilize the fixed alignment in EPD but instead use the time-delay architecture that allows the alignment to vary and appropriate modeling of the two main signals in the core promoter region.

The two time-delay neural networks were trained independently. It was experimentally determined that a receptive field size of 15 bp performed the best. For the TATA network, this leads to a total of 120 input units (30 bp) and 60 weights (4 x 15) for each unit in the hidden layer. The Inr network has 100 input units (25 bp) and also 60 weights (4 x 15) for each unit in the hidden layer.

The weights of the receptive fields for both of the two networks were initialized using the weight matrices from the literature to "push" them to recognize particular signals. The TATA box weight matrix was taken from Bucher (1990), and the Inr weight matrix from Penotti (1990). These initializations were ideal to train the TDNNs to recognize the appropriate signal in the sequence (i.e. the TATA box time-delay network was forced to train only on the TATA box pattern at approximately -20 bp). The results of both networks can be seen in Table 3-1 and are discussed below.

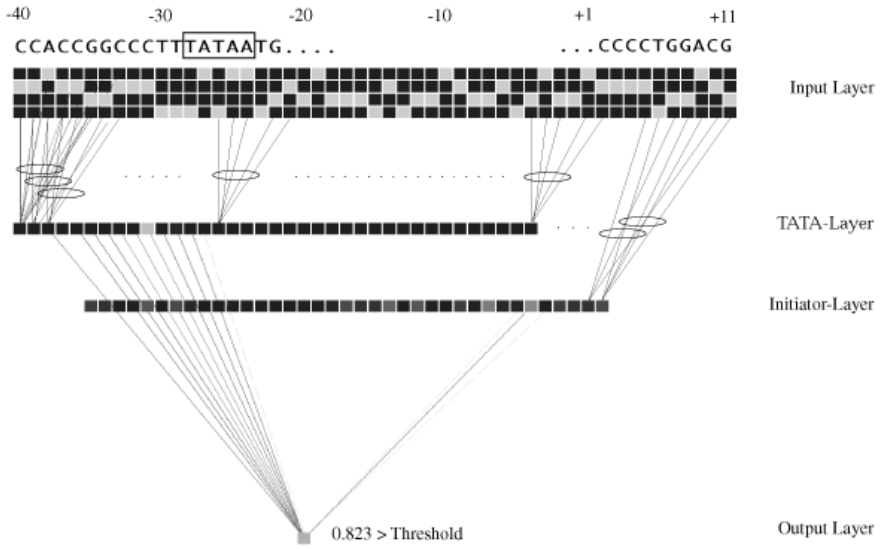
2.6.1 Incorporation of feature detector networks into the final TDNN

To combine the individual feature detector neural networks for TATA and Inr, we use a two-layer time-delay neural network. The input to the final TDNN consists of 51 bp, spanning the transcription start site from position -40 to +11 and including the TATA box and the Inr. The hidden layers from the two previously trained single-feature time-delay neural networks are copied into the combined TDNN and training is carried out. The resulting neural network maps high order correlation between the different features and their relative distance into a complex weight matrix. A snapshot of the two-layer (TATA and initiator) trained TDNN is shown in Figure 2-2. The weights from the hidden layers can be interpreted as the preferred position for an individual element in the input window.

All neural networks were integrated and tested using the Stuttgart Neural Network Simulator Software toolkit (Zell & al., 1999). The networks were then implemented in the **N**eural **N**etwork for **P**romoter **P**rediction (NNPP) program. This program is publicly accessible through a World Wide Web server (http://www.fruitfly.org/seq_tools/promoter.html). It has also been distributed as a

stand-alone program on request and is integrated into the Genotator program (Harris, 1997) as a part of the back-end analysis package.

Figure 2-2: The trained two-layer time-delay neural network. The small squared boxes symbolize the neurons. The input layer is on top with the window reading in the DNA sequence. The receptive fields indicated with a circle grouping connections from the input layer to the two hidden layers (TATA and Inr) show the structure of the time-delay connections. Both hidden layers connect to the single output neuron on the bottom. For clarity, only strong weights are shown. For example, the only significant weights shown from the TATA-layer to the output unit are the ones that localize the position of the TATA box at the beginning of the input window (below CCACCGG). The TATA box is boxed. This test sequence of CCACC….GGACG received a score of 0.823 from NNPP.



Chapter 3 Results of NNPP

This chapter covers the testing of the NNPP program, addresses some of the strength and weaknesses, and gives some examples of its application. The first section (**3.1**) reports the accuracy of NNPP on an assembled collection of eukaryotic promoters. Section **3.2** describes the adaptation of NNPP for contiguous genomic sequence. Sections **3.3** - **3.5** describe the application and give results of NNPP to identify TSS's in various long genomic sequence datasets giving examples from human DNA, from a mixture collection of experimentally verified eukaryotic promoters, published in a comparison in 1997 and, finally, from the *Adh* region in *Drosophila melanogaster*.

3.1 Accuracy of NNPP on a selected promoter dataset

Table 3-1 shows the prediction results for the two single feature time-delay neural networks, the TATA box feature detector (column 2), the Inr feature detector (column 3) and the two-layer TDNN, which incorporates both (column 4 and 5). The results are averaged over four cross-validated test sets produced from the complete dataset of 429 promoters. The correlation coefficient is calculated as defined originally by Matthews (1975) and later adapted to the problem of gene finding evaluation by Burset and Guigó (1996) as:

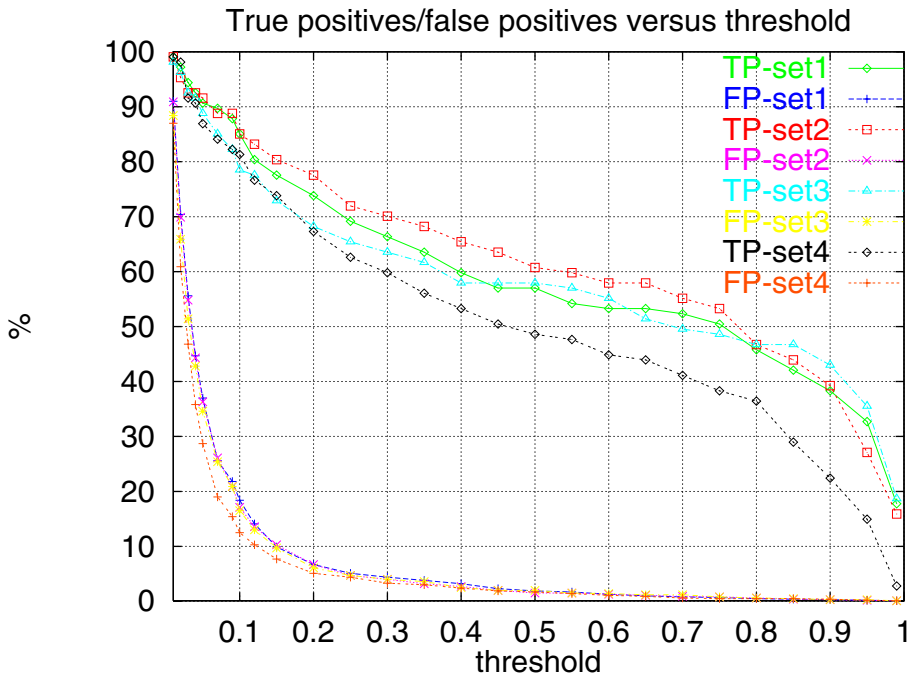
$$CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

Figure 3-1 shows the false positive (FP) and correct positive (CP) prediction results versus the threshold applied to the neural network score. The curves for the individual test sets are almost identical indicating that the sets are independent.

Table 3-1: NNPP Prediction performance on the 4-fold cross-validated data set. False positive rates and correlation coefficients are averaged over the 4-cross validated sets.

False Positive (FP) rates and Correlation Coefficients (CC)					
% Promoters recognized	**TATA box FP-rate (CC)**	**Initiator FP-rate (CC)**	**Combined 2-layer TDNN (CC)**	**Threshold (0-1) for combined TDNN**	**Multi-layer Perceptron FP-rate (CC)**
10	0.2% (0.36)	0.8% (0.28)	0.0% (0.38)	0.99	0.2% (0.35)
20	0.3% (0.45)	2.7% (0.27)	0.1% (0.38)	0.97	0.3% (0.45)
30	0.5% (0.52)	7.0% (0.28)	0.3% (0.50)	0.92	0.8% (0.48)
40	0.9% (0.56)	10.6% (0.26)	0.4% (0.60)	0.85	1.9% (0.50)
50	1.3% (0.62)	18.7% (0.25)	1.0% (0.65)	0.70	3.7% (0.51)
60	3.8% (0.60)	33.0% (0.21)	3.1% (0.61)	0.38	9.9% (0.44)
70	7.2% (0.57)	45.5% (0.18)	5.3% (0.58)	0.20	16.1% (0.40)
80	22.3% (0.39)	60.5% (0.17)	12.5% (0.52)	0.12	45.5% (0.23)

Figure 3-1: Results of the 4-cross validated test sets. Each point on the curves gives the prediction accuracy for a fixed threshold. The four upper curves are the correct positive prediction rates and the lower curves, the corresponding false positive prediction rates for all four test sets.



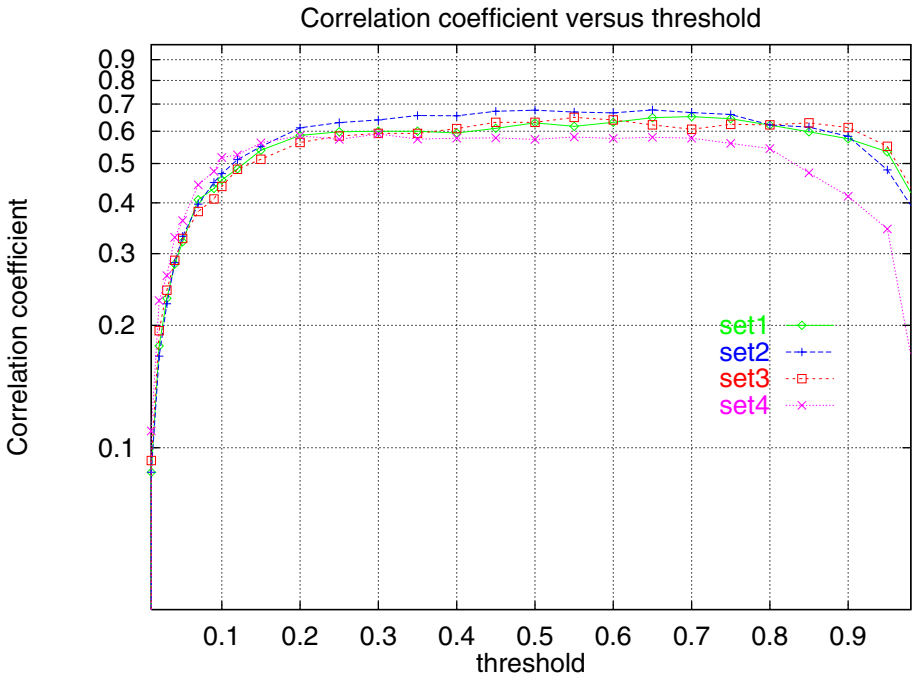
As can be seen from Table 3-1, the performance of the feature detecting networks used in isolation is rather poor. The TATA box network has the better performance of the two, since over 60% of the vertebrate promoters contain a TATA box. The predictive power of the initiator network is weaker because there is no real consensus sequence for vertebrate Inrs. The TATA box network recognizes on average 64 (60%) of the 107 promoter sequences in each test set (4-fold cross-validated) with an average of 38 (3.8%) false positive predictions. If we adjust the threshold so that on average 75 (70%) of the promoters are predicted correctly, there are 72 (7.2%) false positive predictions. The Inr neural network can only detect 11 (10%) of the promoters, with a false positive rate of 0.8%. The combination of both neural networks increases the prediction rate. If on average in the 4 cross-validated sets 54 (50%) promoters are correctly predicted, the false positive rate drops down to 1.0% (ten coding DNA regions predicted as promoters; correlation coefficient of 0.65), but that is similar to the TAT A-only results. Even if 75 (70%) promoters are correctly predicted, the average number of false predictions is only 53 (versus 72 for TATA alone). At a threshold of 0.12, 80% of the promoters predicted, the number of false positive predictions goes up to 125

(12.5%). 21 (19.6%) promoter sites on average in the test sets cannot be predicted at all using this 2-layer neural network.

For comparison, the results for a "standard" feed-forward backpropagation neural network with one hidden layer trained on the same data sets are shown in the last column of Table 3-1. The number of hidden units and the number of training cycles were optimized exactly the same way as for the time-delay neural network. The results show the superiority of the two-layer TDNN. At a threshold that gives 64 (60%) correct predictions, the number of false positive predictions is more than three times higher for the standard network (99 false predictions) than for the 2-layer TDNN (31 false predictions). This shows that reducing the parameter space from 3,091 adjustable weights in the standard network to 169 in the TDNN, improves the prediction accuracy on a limited training data set (419 promoter sequences).

Figure 3-2 shows that the correlation coefficient performance for the 2-layer neural network on all four data sets is dependent on the threshold. The prediction accuracy expressed in the correlation coefficient (CC) gets the highest value, on average, with a threshold of 0.5 and is fairly stable in the range of thresholds 0.2 - 0.9.

Figure 3-2: Correlation coefficient results for the 4-cross validated test sets.



3.2 Application of NNPP to long, contiguous genomic DNA

To apply the 2-layer time-delay neural network to contiguous genomic sequence, a window of 51 basepairs is shifted over the sequence base by base. In this way, a score is computed for every position in the sequence. These individual scores are subsequently smoothed by a simple but efficient function, which selects the position of the highest score in a window of 10 neighboring positions as the final prediction. The smoothing function is implemented as a post-processing procedure and is part of the final NNPP.

3.3 Accuracy of NNPP in human DNA

In Figure 3-3, the output of the smoothing function is plotted for the 2-layer TDNN neural network output for the genomic sequence of a human tissue factor gene (GenBank accession HUMTFPB). In this sequence, NNPP finds the annotated promoter at position 799 with a score of 0.997. With the threshold cut-off used in Figure 3-3 of 0.96, 8 false positives are predicted in the 13,865 bases sequence (forward strand only). This corresponds to a false positive recognition rate of approximately six in 10,000 bases or 0.0577% (forward strand only!). The figure also illustrates the usefulness of the scores. If one had chosen a threshold of 0.99, one would still have found the actual promoter and predicted only two false positives sites (one of them, the second highest scoring at 12,500 in Figure 3-3 might even be a real promoter belonging to the next gene). Another human test sequence is the well-studied gene cluster of growth factors (GenBank accession: HUMGHCSA). In Figure 3-4 the 66,495 bases sequence shows twenty false predicted sites (three in 10,000 basepairs or 0.0301% (forward strand only)). With a threshold of 0.99, all five known promoters are predicted and only one false positive site in the 5' untranslated region of the first gene is over-predicted. These two examples show the variation of the false positive prediction rate in different human genomic sequence regions.

Figure 3-3: NNPP predicted positions for the TSS for the human tissue gene HUMTFPB (13,865 bp). The known coding region including introns (not shown) is indicated as a rectangular box in the top region of the graphs.

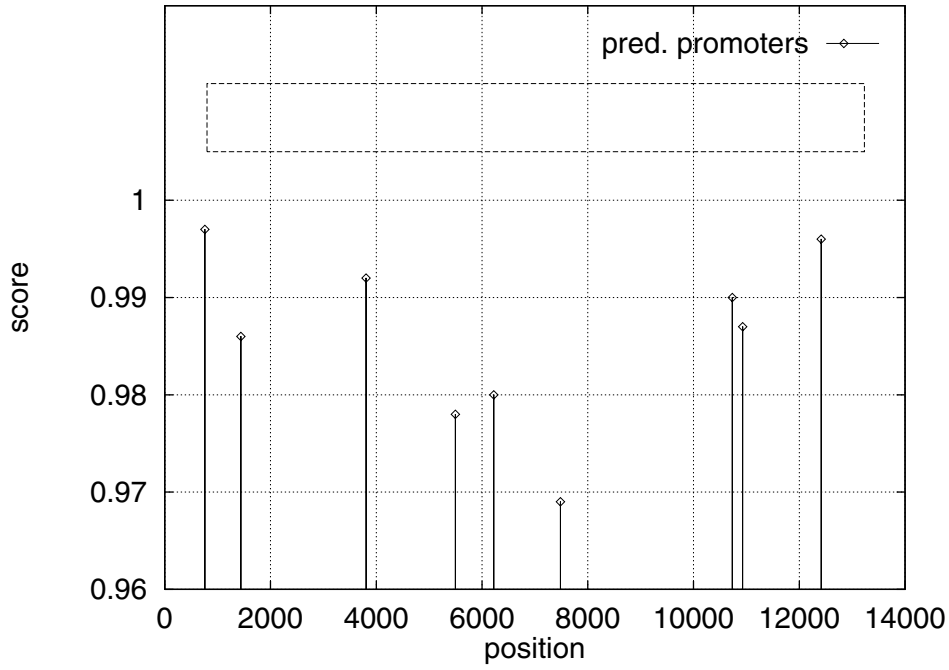


Figure 3-4: NNPP predicted positions for the TSS for the human growth factor region HUMGHCSA (66,495 bp). Known genes are indicated as rectangular boxes in the top region of the graph.

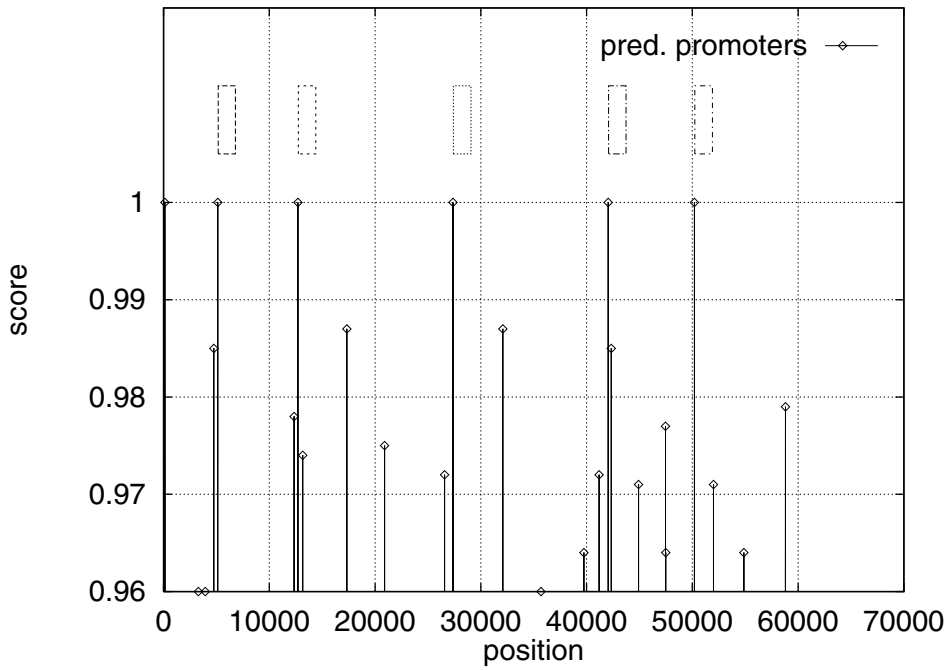


Table 3-2 shows NNPP results on two additional selected human genomic sequence the human beta globin gene (GenBank accession: HUMHBB) and the human herpes simplex virus region (GenBank accession: HEHSV1SU). While the false positive rate for the human beta globin locus of 1.2/10,000 bases (threshold of 0.99) is similar to the examples in Figure 3-3 and Figure 3-4, the human herpes simplex virus reveals many more false positives (almost eight in 10,000 at a threshold of 0.99). This might have to do with the specific gene content in this region or with previously unannotated genes and their corresponding promoters. Table 3-2 also shows NNPP results in comparison to two other existing promoter prediction programs PromoterScan (Prestridge, 1995) and the promoter subprogram of GRAIL2 (Matis et al., 1995). We adjusted NNPP's thresholds so that the false positive rates are comparable to that of the other programs. While both programs seem to predict fewer false positive sites both also miss many known promoter sites. The lower false positive prediction rates are not too surprising, because PromoterScan uses a long promoter region for scoring and NNPP uses only 51 basepairs. Matis' promoter recognition method, which is integrated into the GRAIL2 gene finding system, also reduces the false positive rate.

Table 3-2: NNPP results in human DNA. Results obtained with the genomic sequence of the human herpes simplex virus (HEHSV1SU: nine genes and eleven TSS's (two genes with two alternative TSS's)) the human tissue gene (HUMTFPB: one gene and one TSS), the human beta globin region (HUMHBB: six genes and seven TSS's (one gene with two alternative TSS's)), and the human growth factor region (HUMGHCSA: five genes and five TSS's) using NNPP, PromoterScan version 1.5 and Grail 2 version 1.3b. We show two result sets (threshold 0.8 and 0.99) for our method. For the two other methods the default settings were used.

	NNPP (t=0.8)		NNPP (t=0.99)		PromoterScan		GRAIL2	
	CP	**FP**	**CP**	**FP**	**CP**	**FP**	**CP**	**FP**
HEHSV1SU Herpes simplex virus (12,979 bp)	10/11	0.39%	7/11	0.077%	3/11	0.054%	2/11	0.015%
HUMTFPB Human tissue gene (13,865 bp)	1/1	0.19%	1/1	0.014%	1/1	0.0072%	1/1	0.0072%
HUMHBB Human beta globin gene (73,308 bp)	7/7	0.13%	1/7	0.012%	2/7	0.0082%	2/7	0.0082%
HUMGHCSA Human growth factor cluster (66,495 bp)	5/5	0.12%	5/5	0.0015%	4/5	0.014%	5/5	0.011%

3.4 Accuracy of NNPP in a eukaryotic promoter recognition assessment project

In 1997, Fickett and Hatzigeorgiou (1997) published a first overview of the status of eukaryotic promoter recognition algorithms. They compared the computer methods "Audic" (Audic & Claverie, 1997), "Autogene" (Kondrakhin *et al.*, 1995), "GeneID/Promoter1.0" (an older version of Promoter2.0 (Knudsen, 1999)), "NNPP" (through the web interface), "PromFind" (Hutchinson, 1996), "PromoterScan" (Prestridge, 1995), "TATA" (derived from (Bucher, 1990)), "TSSG" and "TSSW" (Solovyev & Salamov, 1997).

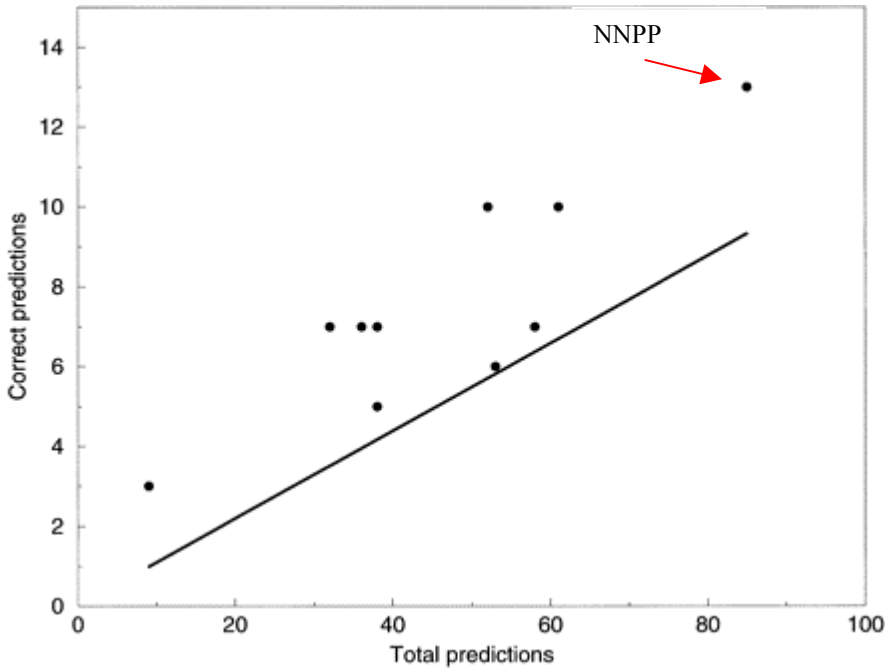
All eight programs were tested on eighteen published mammalian sequences containing twenty-four promoters (see Table 1 in (Fickett & Hatzigeorgiou, 1997) for a list) in a total of 33,120 basepairs of genomic DNA. They found recognition rates, expressed as sensitivity, on the order of 13%-54% of the true promoters and false positive rates on the order of 1/460 to 1/5,520 (see Table 3-3 from the original Table 2 in the Fickett and Hatzigeorgiou publication (1997)).

Table 3-3: Comparison of performance accuracies. Taken from (Fickett & Hatzigeorgiou, 1997). Program Accuracy. Overall accuracy of the programs tested. For each program the sensitivity (both as the number and percentage of promoters correctly detected) and specificity (as number of false positives and number of basepairs per false positive) is given.

	Audic	**Auto gene**	**Gene ID**	**NNPP**	**P'Find**	**P'Scan**	**TATA**	**TSSG**	**TSSW**
Sensitivity	5/24 24%	7/24 29%	10/24 42%	13/24 54%	7/27 29%	3/24 13%	6/24 25%	7/24 25%	10/24 42%
Specificity	33 FP 1/1004 bp	51 FP 1/649 bp	51 FP 1/649 bp	72 FP 1/460 bp	29 FP 1/1142 bp	6 FP 1/5520 bp	47 FP 1/705 bp	25 FP 1/1325 bp	42 FP 1/789 bp

In addition to Table 3-3 the accuracy of the various programs are plotted in Figure 3-5 (also taken from the original Figure 1 in the Fickett and Hatzigeorgiou (1997) publication). This figure combines the accuracy measures of sensitivity and specificity from Table 3-3. The accuracies of the programs whose results are furthest away from the plotted random prediction axis are the highest. As one can see, NNPP's (marked in Figure 3-3) is one of the best performers. The figure also shows that NNPP has a high sensitivity but a very low specificity (see Table 3-3 for exact scores). This is due to the small window on which the prediction is based. The NNPP approach falls into the category of "search by signal" approaches (see Haussler (1998)) and has the highest coverage of real TSS's (54%). Most of the other programs, for example Audic (Audic & Claverie, 1997), PromFind (Hutchinson, 1996) and TSSG/TSSW (Solovyev & Salamov, 1997), fall in the "search by content" category, because they take into account general DNA statistics such as coding potential, nucleotide frequencies, such as CpG islands, AT-richness and others. PromoterScan (Prestridge, 1995) is a "search by signal" method that uses heuristics to combine several weight matrix hits. All these programs make fewer false positive predictions but miss more true TSS's.

Figure 3-5: Taken from (Fickett & Hatzigeorgiou, 1997). Each point plotted represents the accuracy of one program, with the abscissa being the total number of predictions made by the program, and the ordinate being the number of correct predictions. For comparison, the line y = 0.11x is plotted. 0.11 is the fraction of all bases in the test set where a prediction would be counted as correct, so that points on the line would reflect the accuracy, on average, of random predictions.



To use content information for a better prediction in section 4.4.3 below I describe the integration of NNPP into the Genie gene finding system. This system then integrates "search by signal" and "search by content" methods into a recently defined third category of promoter prediction programs called "promoter prediction through gene finding" (Reese et al., 2000).

3.5 Application of NNPP in *Drosophila melanogaster*: The *Adh* region

To test the accuracy of NNPP in *Drosophila melanogaster*, NNPP was applied to the 2.9 Mb genomic sequence of the *Adh* region (Ashburner et al., 1999). As part of the GASP genome annotation assessment experiment (Reese *et al.*, 1999), Uwe Ohler prepared a set of annotations of very likely TSS's for the entire *Adh* region. He describes in Reese *et al.* (2000) that prior to the experiment almost no experimentally confirmed annotation of a TSS existed in the *Adh* region. To extend the very few experimentally confirmed TSS's, he used gene annotations from Ashburner *et al.* and specifically the 5'

UTR alignments of the existing full-length cDNA sequences and 5' EST clusters to obtain the best approximation for each TSS. Because 5' UTRs of *Drosophila* genes can extend up to several kilobases upstream of the ATG, this was not an easy task. A careful analysis resulted in high quality full-length cDNA alignments for 92 genes out of the original 222 gene annotations. Full-length cDNA sequences were taken from recent cDNA sequences, described in Reese *et al.* as the *std1* set, as well as previously known cDNA sequences from GenBank (for more details see (Reese et al., 2000)). This recent experiment represents the first assessment of promoter prediction techniques for a significant number of genes in a large contiguous genomic region.

In Table 3-4 the NNPP results are reported on this test set of genes in the *Adh* region (Ashburner et al., 1999) in comparison to CoreInspector (Scherf *et al*., 2000) and MCPromoter (Ohler et al., 1999) in the GASP study (Reese et al., 2000), excluding the "promoter prediction through gene finding" programs. Although NNPP is far from accurate, this test shows results similar to those in the 1997 review by Fickett and Hatzigeorgiou (see Table 3-3). They reported a recognition rate of 54% of the known promoters at a threshold of 0.8. In *Adh*, the same threshold identifies 69 or 75% of the total of 92 annotated promoters with a false positive rate of 1/547, similar to the rate of 1/460 reported in (Fickett & Hatzigeorgiou, 1997). It has to be noted that Fickett and Hatzigeorgiou used both strands to calculate the false positive rate while for *Adh* only the gene strand was used. If one applies a more stringent threshold of 0.97, 35 of the 92 promoters are still recognized with a much lower false positive rate of 1/2,416. The higher classification rate might be due to biased promoter selection in (Fickett & Hatzigeorgiou, 1997). Compared to the human sequences in Table 3-2 the false positive results on the *Drosophila* genome DNA are a little higher (t=0.99: 0.013% (human) vs. 0.016% (*Drosophila*)).

Table 3-4: Evaluation of promoter prediction systems on the *Adh* region. The table only shows the results of the "search by signal" program (CoreInspector) and "search by content" programs (MCPromoter) from the experiment of Reese *et al.* (2000) and the prediction sets from NNPP with different thresholds. Only the identified TSS's from *Adh* with long cDNA alignments are shown (for a full explanation of the test data set, see Reese *et al.* (2000)). The rate of false positives is shown for the sequence where cDNA annotations define the region as non-promoter.

	System name	Identified TSS	Rate of false predictions in annotated *Adh* region (total 853,180 bases)
From Reese et al., 2000	**CoreInspector**	1 (1.0%)	1/853,180 (0.00012%)
	MCPromoter v2.0	31 (33.6%)	1/2,437 (0.041%)
	MCPromoter v1.1	26 (28.2%)	1/2,633 (0.038%)
Plain NNPP	**NNPP (t=0.99)**	20 (21.7%)	1/6,227 (0.016%)
	NNPP (t=0.97)	35 (38.0%)	1/2,416 (0.041%)
	NNPP (t=0.92)	49 (53.2%)	1/1,096 (0.091%)
	NNPP (t=0.90)	55 (59.7%)	1/928 (0.108%)
	NNPP (t=0.85)	65 (70.6%)	1/685 (0.146%)
	NNPP (t=0.80)	69 (75.0%)	1/547 (0.183%)
	NNPP (t=0.70)	80 (86.9%)	1/400 (0.250%)
	NNPP (t=0.38)	91 (98.9%)	1/164 (0.610%)
	NNPP (t=0.20)	92 (100.0%)	1/75 (1.333%)

The statistics formulated in Fickett and Hatzigeorgiou (1997) where based on a window of -200 to +100 bases centered around the TSS while in the *Adh* GASP experiment a window of +500 to -50 bases for a correct prediction around the translation start site was used in absence of an annotated TSS. This is very generous. A check of the exact position of the TSS predictions using a threshold of 0.97 for NNPP shows that for the 35 predicted TSS's, the average distance from the predicted to the annotated TSS is 148.94 bases. This is surprisingly high and might be due to TSS annotation errors in the *Adh* standard set. The methods of aligning cDNAs to genomic DNA to identify the TSS are known to be problematic. Predictions from MCPromoter and NNPP (t=0.97) agree in 11 out of 35 cases but again the exact predicted positions do not show a strong correlation. The average distance from the predicted TSS to the annotated TSS for MCPromoter is 131.24, which is more precise than NNPP.

Chapter 4 Gene finding using a generalized hidden Markov model (Genie)

In this chapter, the structural and compositional features of genes in genomic DNA are presented and a probabilistic model of gene structure is introduced. Section 4.1 reviews the biological features of a gene and summarizes historical and current approaches to computational gene finding. Section 4.2 presents the data sets of known *Drosophila* genes and the genomic region, which is used for evaluating Genie, is introduced. In Section 4.3 the basic HMM framework for the gene model is described, and a description of the individual submodels used in the overall framework is given. Section 4.4 discusses the implementation of Genie and describes the three versions of Genie: Genie, GenieEST and GenieESTHOM. Finally the algorithmic integration of the time-delay neural network into the Genie system is presented.

4.1 Background

Recent advances in sequencing technology are making the generation of whole genome sequences commonplace. Immediately after deciphering the nucleotides of a genome, the process of interpretation of these raw data into useful biological information begins. The first features in a genome to detect and describe are the genes. The major class of genes is the class of protein coding genes. Traditionally, small-scale studies of isolated genes were carried out in an individual researcher's laboratory. They used a combination of computational and experimental methods that permit very detailed description of the gene and its features. In contrast, for complete genomes robust automatic methods are needed to identify the majority of the genes quickly.

In prokaryotes, and in some simple eukaryotes such as *Saccharomyces cerevisiae*, genes normally have single continuous open reading frames (ORFs) separated by short intergenic regions. In contrast genes in most eukaryotes may be very complex, with many exons, introns that may be ten's of kilobases in length, non-coding 5' and 3' exons and alternatively spliced products. In addition complex relationships between genes may be quite frequent, e.g. genes contained within the introns of other genes and adjacent series of very related genes. Therefore, methods for these higher organisms (including *Drosophila*) have to be much more complex and are much less robust and sensitive. A recent study in *C. elegans*, for example, has shown that the first path annotation process of genes can provide an overview of the entire genome, but that it is rather superficial and incomplete in describing individual genes. In this particular genome, the originally estimated number of total genes of approximately 19,000 had to be reduced to 14,000 - 15,000. This also shows the dependence of genome annotations on the tools used especially the gene identification systems.

Besides giving an accurate estimate of the existing genes in a first path annotation phase, any improvement in the accuracy of the predicted gene structure will be of high value for subsequent genome analysis because error correction through biological experiments is very labor intensive and expensive. Complete correct predictions cannot be produced by the current gene finding technologies. Therefore the *Drosophila* Genie system has been evaluated extensively by participating in the international assessment project called the Genome Annotation Assessment Project (GASP) organized by members of the BDGP (Rubin & al., 1999) including myself (Reese et al., 2000). This experiment was a blind test and the assessment was based on the *Adh* region (Ashburner et al., 1999). This assessment will provide biologists that rely on computational annotations for biological studies with confidence values for the individual predictions, which is a necessity to understand the value and quality of any annotation of a genome.

In this chapter, I will briefly describe the general task of identifying genes in genomic DNA, introduce the methodology of a generalized hidden Markov models, discuss the implementation of the computer system and the training on *Drosophila* genes, discuss shortly the evaluation process used at GASP to better understand the performance results, report on results of Genie on the *Adh* region (as evaluated at GASP) and give a short report of the application of the final Genie system to the entire genome of *Drosophila melanogaster*.

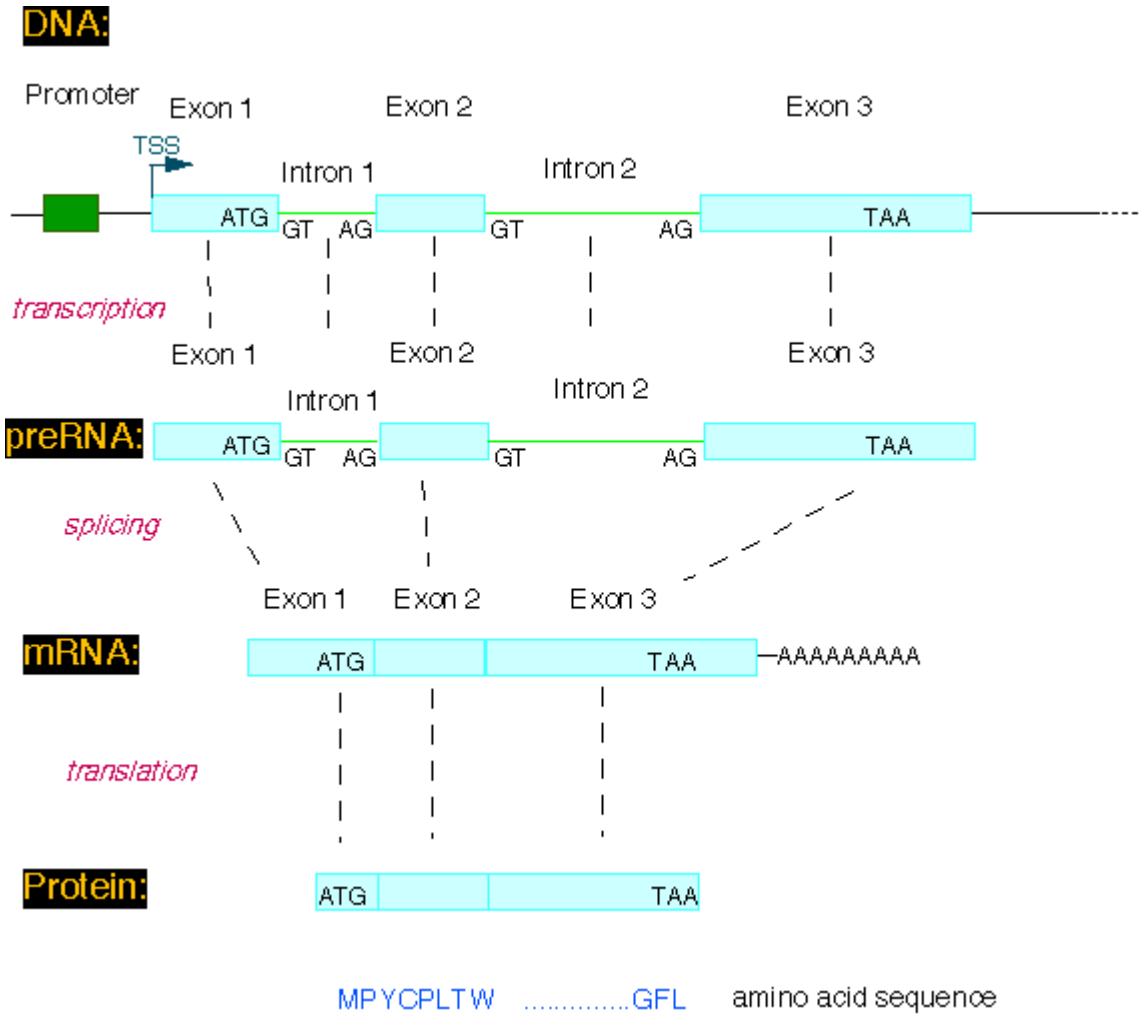
4.1.1 The structure of *Drosophila* genes

Genes in *Drosophila* have a complex genomic structure. A schematic can be seen in Figure 4-1. In a very top down view a gene consists of exons and introns and the two transcription regulatory sites, the promoter and the poly-adenylation site. Promoters have been discussed in detail in the previous chapter and for poly-adenylation sites I refer to the textbook literature because they have not played a major role in the development of the Genie system owing to their lower statistical significance. The transcription machinery initiated at the TSS, transcribes the entire gene up to the poly-adenylation site into pre-mRNA. Immediately afterwards the exons in the pre-mRNA are merged together through the splicing process. This splicing process of the removal of the introns is carried out in the cell nucleus by a complex process catalyzed by a 60S particle known as the spliceosome. The spliceosome is composed of five small nuclear RNAs (snRNAs) called U1, U2, U4, U5 and U6, and numerous other proteins (for an overview see (Neubauer *et al*., 1998)). Splice site recognition and spliceosomal assembly occur simultaneously according to a complex sequence of steps (for the most comprehensive review see (Green, 1991; Moore *et al*., 1993)). The consensus sequence for the 5' splice site (exon-intron) contains the conserved motif "GT" following the exon and the 3' splice site (intron-exon) contains the conserved motif "AG" preceding the exon. Exceptions for these conserved motifs are the so-called U2 or U12 type introns (Sharp & Burge, 1997) mostly found and studied in human. The 5' and 3' splicing signals are statistically the strongest signal for gene recognition.

After splicing, the mRNA leaves the nucleus and gets translated at the ribosome into the amino acid chain. The genetic code determines the translation from the nucleotides into the amino acid sequence. The typical usage of an organism of a certain set of codons is called codon preference. It has strong statistical features and varies among organisms and species. The statistical preferences for a protein sequence are also organism specific. This is usually called codon usage. Both features are statistically significant and are therefore the most contributing features models in the computational models. The translated region starts with a methionine (M), which is coded as AUG in the mRNA and ATG in the genomic sequence. The end of translation is marked by stop codons. In *Drosophila* these are TAA, TTG and TTA.

Additional gene features such as the branch point, transcription factor binding sites, leader or signal peptides, enhancer sites or other regulatory sites are not modeled in Genie owing to the low statistical significance for a general gene model and our limited understanding of the underlying processes.

Figure 4-1: Schematic gene structure.



4.1.2 Computational gene finding

Computational gene finding can be divided into two major classes of techniques for the prediction of genes - *ab initio* methods and homology-based methods. Given the complexity of eukaryotic genes described above any *ab initio* method must combine the prediction of gene components, exons, introns, splice sites etc, with the prediction of a model of how these components may be assembled into a gene. Homology-based

methods on the other hand can rely on information from biological experiments, cDNAs or proteins from other organisms to cope with these gene assembly difficulties.

Computational gene finding has evolved steadily over the last 20 years and excellent reviews in this area have been written by Fickett (1992), Claverie (1997), Guigó (1997) and Burge (1998). In the most recent review, Haussler (1998) has categorized submodels in gene finding methods as either "signal sensors" or "content sensors". In broad terms signal sensor methods exploit descriptions of pertinent sites such as splice junctions, start and stop codons, branch points, promoters, termination of transcription and others to identify genes. Content sensor methods employ models that are based upon extended lengths of sequence such as exons and introns.

Pioneering studies in eukaryotic gene identification (Fickett, 1982; McLachlan *et al*., 1984; Shepherd, 1981; Staden & McLachlan, 1982; Trifonov & Sussman, 1980) showed that statistical measures related to biases in codon preference and codon usage could be used to identify protein-coding regions. Since then many methods have been developed that show differences in coding versus non-coding genomic sequence including *k*-tuple frequencies (Claverie & Bougueleret, 1986), measures of auto-correlation (Michel, 1986), spectral analysis using Fourier transformation (Silverman & Linsker, 1986) and G-notG-U periodicity statistics (Trifonov, 1987). Based on these statistical differences a first generation of computer programs was developed to identify approximate coding regions in genomic DNA sequence. The most well known programs are TestCode based on Fickett's work (1982) and GRAIL (Uberbacher & Mural, 1991), which is based on a neural network approach that integrates multiple gene features of content type (i.e. exons) to classify genes. Generally these first generation programs were able to identify the approximate location of coding regions.

The next wave of programs addressed the need more precisely to determine the exon boundaries. These programs, such as GRAIL II (Xu *et al*., 1994a) and Xpound (Thomas & Skolnick, 1994), used *ad-hoc* methods to integrate splice site signals and content measures for refining exon prediction. The third class of gene finding systems attempted the task of assembling separate exon predictions into a complete gene structure, which would predict the entire protein sequence. The first systems were *gm* (Fields & Soderlund, 1990) for *C. elegans* genes and a mammalian system by Gelfand (1990). In an excellent review by Fickett and Tung (1992) of the different coding versus

non-coding potential methods, various different approaches to complete gene prediction were presented: GeneID (Guigo *et al*., 1992), a fast hierarchical ruled based system, GeneParser (Snyder & Stormo, 1993; Snyder & Stormo, 1995; Stormo & Haussler, 1994), a neural network approach combined with pioneering work on dynamic programming for this problem, GenLang (Snyder & Stormo, 1993; Snyder & Stormo, 1995; Stormo & Haussler, 1994), an approach based on linguistic state machines, Fgene (Solovyev *et al*., 1995), a linear discriminant analysis included in dynamic programming system, and the GRAIL based system GAP (Xu *et al*., 1994b).

For dynamic programming methods the key to success is developing the right scoring function to optimize. A fruitful approach here has been to define a statistical model of genes that includes parameters describing codon dependencies in exons, characteristics in splice sites and other signals as well as a "state machine" containing information on what functional components are likely to follow others. The earliest theoretical work by (Stormo & Haussler, 1994) has inspired many followers to explore the power of a statistical framework for a gene model. The "state machine" conceptual structure has mostly been expressed by the parameters of a Markov process on the hidden variables. Therefore they are called hidden Markov models and can be seen as stochastic grammar models. Gene finding systems of this fourth generation are: an extended version of Xpound (Thomas & Skolnick, 1994), GeneMark.hmm (Lukashin & Borodovsky, 1998), a system based on the original prokaryotic version of the program, Veil (Henderson *et al*., 1997), and HMMGene (Krogh, 1997), a system conceptually based on the very early HMM work in *E.coli* called EcoParser (Krogh *et al*., 1994b). A more general class of probabilistic model, called a generalized HMM (GHMM), which is the closest implementation of the theoretical work by Haussler and Stormo (1994) and sometimes also called semi-hidden Markov model, was fully developed in the earliest Genie implementation (Kulp *et al*., 1996; Reese *et al*., 1997) and subsequently by GENSCAN (Burge & Karlin, 1997), which uses different training sets and different implementations of the so-called submodels.

A good state-of-the-art performance overview gives the work by Burset and Guigó (1996) on human DNA. For *Drosophila* the GASP experiment presented at the 7th International conference on Intelligent Systems in Molecular Biology computational gives an excellent assessment and state-of-the-art overview of the technology. GASP was a blind test and was performed on a well-studied 2.9 mega-base sequence region of

the *Drosophila melanogaster* genome (Ashburner et al., 1999) (further described below).

4.2 Gene datasets: Training of Genie

A clear outcome from the GASP experiment (Reese et al., 1999) is that clean annotations from well-studied regions in genomes are absolutely essential effectively to evaluate, compare, and refine existing annotation methods. It is likewise clear that they are equally essential for improved, curated training sets used to train the models for gene prediction. Therefore I will discuss briefly the test bed for my studies, the genomic *Adh* region, as well as giving a short overview of the process that went into creating my training sets.

4.2.1 Genomic DNA sequence

The selection of a genomic target region for assessing the accuracy of computational genome annotation methods is a difficult task for several reasons: The genomic region has to be large enough, the organism has to be well studied, and enough auxiliary data has to be available to have a good experimentally verified "correct answer" but the data should be anonymous so that a blind test is possible. The *Adh* region of the *Drosophila melanogaster* genome met these criteria. *Drosophila melanogaster* is one of the most important model organisms and especially the *Adh* region had been extensively studied genetically together with other experiments such as cDNA sequencing.

The 2.9 megabase *Adh* contig is large enough to be challenging, contained genes with a variety of sizes and structures, and included regions of high and low gene density. The test was not a completely blind test, however, since several cDNA and genomic sequences for known genes in the region were available prior to the experiment. A handful of genes were also included in some parts of the Genie training procedure. The annotation for the entire contig has recently been released (Ashburner et al., 1999) and is the basis for the evaluation of Genie and GASP.

4.2.2 Curated training sequences

In order to derive a realistic description of the structural and compositional features of *Drosophila* genes, large non-redundant sets of sequences are required. Especially in

the case of complex models with many parameters to optimize, the requirement for very clean datasets is extremely important. Gene models such as Genie belong to this class.

A collection of 275 multiple exon (Appendix C) and 141 single exon (Appendix D) non-redundant genes together with their start and stop codons and splice site structures in genomic DNA was compiled. This set was constructed by searching the GenBank nucleotide acid sequence database (version 109) for sequences containing single complete *Drosophila* genes (i.e. containing at least the initial ATG through to the stop codon) sequenced on the genomic level as opposed to the mRNAs. Certain conditional constraints were imposed on the data to filter out noisy and questionable entries: Only one CDS entry was allowed, which avoided alternative spliced genes; the annotation should be minimally self-consistent (no inframe stop codon and the minimal consensi for start codon, 5' and 3' splice sites, GT and AG respectively, should be matched); no pseudo genes were allowed; and genes marked as "putative" or "predicted" were removed.

To retain a representative non-biased data set all related genes were removed from the filtered data where a sequence identity of >=80% using BLAST (Altschul *et al*., 1990) was detected. This data set is publicly available at http://www.fruitfly.org/sequence/drosophila-dataset.html in Genbank flat file format and as separate start codon, 5' splice site and 3' splice site sequence sets. These sets have been used as the basis for the training in the GASP experiment so that a fair comparison of the applied methods was possible because they were all trained on the same underlying data sets.

For the individual training of the coding exon models in Genie, which is not so dependent on high quality annotations, a larger collection of all available mRNA sequences from GenBank and directly from the BDGP site was used.

4.3 A generalized hidden Markov model for gene finding

Hidden Markov models are an extension to discrete-time Markov processes. They have been studied and applied mostly in speech recognition. Rabiner (1989) provided an excellent, detailed tutorial. They have recently begun to be used extensively in the computational biology field. In particular they have been used for amino acid profile

searching. (See the PFAM database of HMM profiles: http://www.sanger.ac.uk/Software/Pfam; (Bateman *et al*., 2000); see also Durbin *et al.* (1998)) Hidden Markov models can be used to model real world signals, whether they are characters from some alphabet, speech pattern or temperature readings. There are three reasons for the power of modeling signals using hidden Markov models:

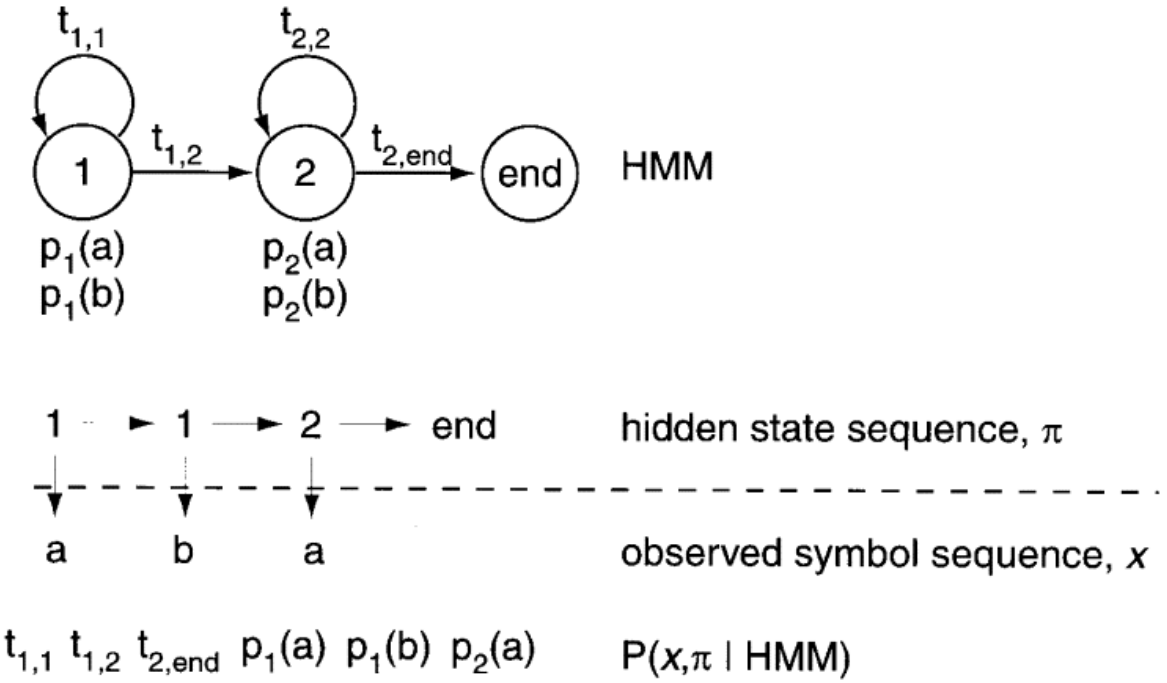
- A model of the signal can be used to derive the theory for the signal-processing program, which in turn will provide useful output from the signals.

- The source of the signal can be studied, even if it cannot be seen.

- Signal models work well in making predictions and recognition.

A hidden Markov model describes a probability distribution over a potentially infinite number of sequences. An example of a simple HMM that models sequences composed of two letters (*a,b*) is shown in Figure 4-2. This toy HMM (taken from Eddy (1998)) would be an appropriate model for a problem in which we thought sequences started with one residue composition (a-rich, perhaps), and then switched once to a different residue composition (b-rich, perhaps). The HMM consists of two states connected by state transitions. Each state has a symbol emission probability distribution for generating (matching) a symbol in the alphabet. It is convenient to think of an HMM as a model that generates sequences. Starting in an initial state, we choose a new state with some transition probability (either staying in state **1** with transition probability $t_{1,1}$, or moving to state **2** with transition probability $t_{1,2}$); then we generate a residue with an emission probability specific to that state (e.g. choosing an a with $p_1(a)$). We repeat the transition/emission process until we reach the end state. At the end of this process, we have a hidden state sequence that we do not observe, and a symbol sequence that we do observe.

Figure 4-2: A toy HMM, modeling sequences of a's and b's as two regions of potentially different residue composition (taken from Eddy (1998)). The model is drawn (top) with circles for states and arrows for state transitions. A possible state sequence generated from the model is shown, followed by a possible symbol sequence. The joint probability P(x,¦HMM) of the symbol sequence and the state sequence is a product of all the transition and emission probabilities. Notice that another state sequence (1-2-2) could have generated the same symbol sequence, though probably with a different total probability. This is the distinction between HMMs and a standard Markov model with nothing to hide: in an HMM, the state sequence (e.g. the biologically meaningful alignment) is not uniquely determined by the observed symbol sequence, but must be inferred probabilistically from it.



The name “hidden Markov model” comes from the fact that the state sequence is a first-order Markov chain, but only the symbol sequence is directly observed. The states of the HMM are often associated with meaningful biological labels, such as “exon position 10”. In our toy HMM, for instance, states **1** and **2** correspond to a biological notion of two sequence regions with differing residue composition. Inferring the alignment of the observed protein or DNA sequence to the hidden state sequence is like labeling the sequence with relevant biological information.

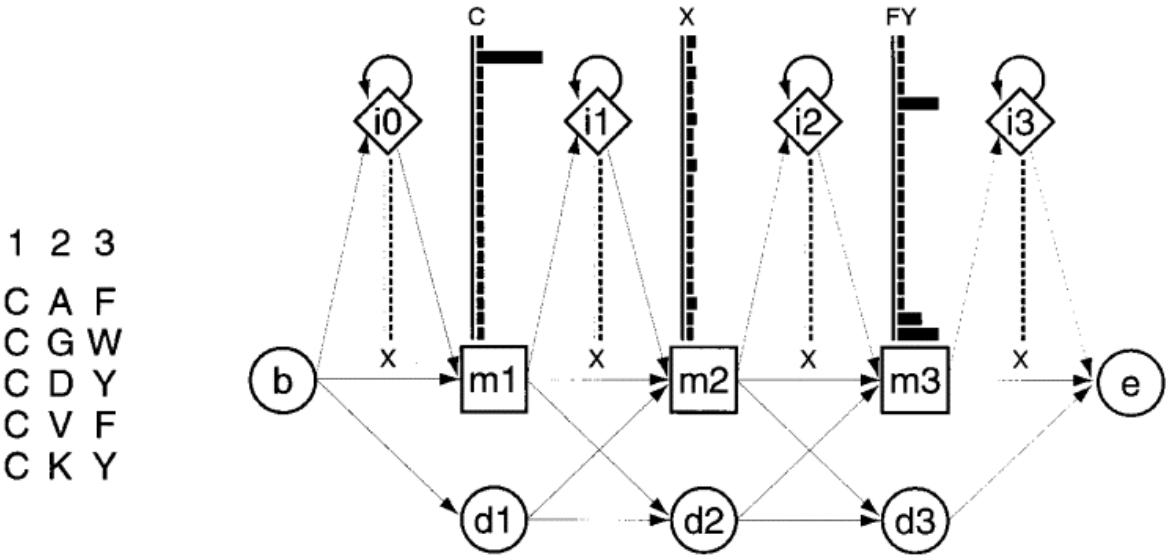
Once an HMM is drawn, regardless of its complexity, the standard dynamic programming local alignment algorithms can be used for aligning and scoring. In the

HMM terminology these algorithms are called Forward (for scoring) and Viterbi (for alignment).

Model Parameters can be set for an HMM in two ways. An HMM can be trained from initially unaligned (unlabeled) sequences. Alternatively, an HMM can be built from pre-aligned (pre-labeled) sequences (i.e. where the state paths are assumed to be known). In the latter case, the parameter estimation problem is simply a matter of converting observed counts of symbol emissions and state transitions into probabilities. Training algorithms are of interest because we may not yet know a plausible alignment for the sequences in question. The standard HMM training algorithms are Baum-Welch expectation maximization (Baum, 1972) and gradient descent.

HMMs were introduced into computational biology in the late 1980s (Churchill, 1989), and for use as so-called “profile” models in 1994 by David Haussler's group at the University of Santa Cruz (Brown *et al*., 1993; Krogh *et al*., 1994a). Krogh *at al.* (1994a) introduced an HMM architecture that was well suited for representing profiles of multiple sequence alignments. For each consensus column of the multiple alignment, a “match” state models the distribution of residues allowed in the column. An “insert” state and “delete” state at each column allow for insertion of one or more residues between that column and the next, or for deleting the consensus residue. Profile HMMs are strongly linear, left-right models, unlike the general HMM case. Figure 4-3 (taken from Eddy (1998)) shows a small profile HMM corresponding to a short multiple sequence alignment.

Figure 4-3: A small profile HMM (right) representing a short multiple alignment of five sequences (left) with three consensus columns (taken from Eddy (1998)). The three columns are modeled by three match states (squares labeled m1, m2 and m3), each of which has 20 residue emission probabilities, shown with black bars. Insert states (diamonds labeled i0-i3) also have 20 emission probabilities each. Delete states (circles labeled d1-d3) are "mute" states that have no emission probabilities. A begin and end state are included (b,e). State transition probabilities are shown as arrows.



Additional background information on HMMs can be found in the excellent reviews of the application of HMMs in molecular biology by Krogh (1998), Eddy (1998) and in Durbin *et al.* (1998).

4.3.1 Generalized hidden Markov models

A generalized hidden Markov model (GHMM) is an enhancement of a standard HMM model as described in the previous section. In a standard hidden Markov model, viewed as a generator, each state emits a single symbol. A GHMM is a more general model in which each state can emit one or more symbols according to an arbitrary distribution. Each state presents an independent submodel, which may itself be a hidden Markov model or any statistical model.

Figure 4-4 shows a simple GHMM that models eukaryotic gene structure. The GHMM is represented as a graph. The states in the model are shown as the arcs of the graph. Nodes in the graph represent transitions between states. (This is different from the typical graphical representation of regular HMMs as shown in the previous section.)

Each state corresponds to a submodel of an abstract gene feature such as an "Internal Exon" (E) or an "Intron" (I). For any sequence of bases, x, and state q, the submodel associated with the state q defines a likelihood for the sequence x. This likelihood is denoted $P(x|q)$. When the GHMM is viewed as a generative statistical model, this is the probability that the sequence x is emitted when the hidden Markov process is in state q. The likelihood functions, one for each state, are part of the definition of the GHMM.

The graph of a GHMM has a unique source node B (for Begin) and a unique sink node F (for Final). The process of generating a string from a GHMM can be viewed as taking a random walk in the graph for the GHMM from the source to the sink. For any state q, the node that the arc for state q leads to, is denoted node (q). Once in this node, a next state is chosen at random from among the outgoing arcs from this node, independent of any previous choices. The probability of choosing the next state r is denoted $P(r|\text{node}(q))$. For example, in Figure 4-4, the state I (Intron) leads to the node (A) (Acceptor). After the acceptor can come either the internal exon state (E) or the final exon (EF). The former is chosen with probability $P(E|A)$ and the latter with probability $P(EF|A)$ where $P(E|A)+P(EF|A) = 1$. These parameters are part of the definition of the GHMM, and are in practice determined from training data, as are parameters defining the likelihood functions $P(x|q)$ defined above.

The full process of generating a string from a GHMM consists of a sequence of random choices: First a state q_1 is chosen from among the outgoing arcs of the source node B. Then a substring x_1 is generated according to the probability distribution $P(\exists|q_1)$. Then a next state q_2 is selected from among the outgoing arcs from node(q_i). Then a substring x_2 is generated according to the probability distribution $P(\exists|q_2)$, etc., continuing like this until a state q_k that leads to the sink node is selected. This state emits the last substring x_k. The full string emitted by the HMM is the concatenation $X = x_1 \ldots x_k$ of all substrings that are emitted. All random choices made in the process of generating the string x are independent, except for the dependencies in the sequence $q_1 \ldots q_k$ of states, which form a Markov chain. In the application of GHMMs, this sequence of states is not observed; only the sequence X is observed.

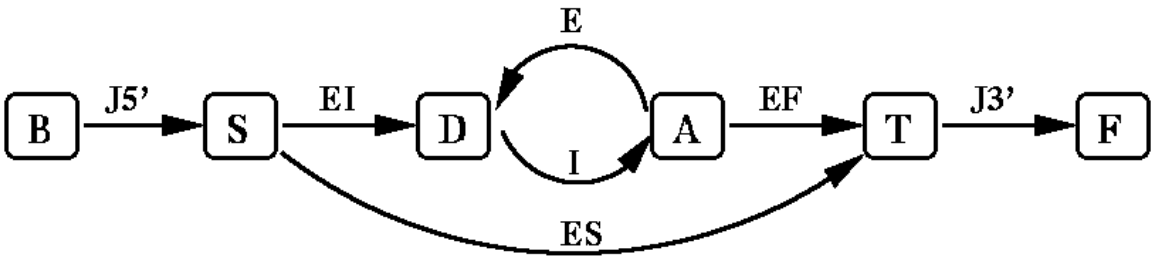
A *parse* Φ of the sequence X is defined to be a pair consisting of a sequence of states $q_1 \ldots q_k$ and a corresponding sequence of substrings $x_1 \ldots x_k$, where $X = x_1 \ldots x_k$, q_1 is a state arc coming out of the unique source node (B), and q_k is a state arc leading to

the unique sink node (F). The GHMM defines a joint likelihood of the sequence $X = = x_1 \ldots x_k$, and the parse $\Phi = (q_1 \ldots q_k; x_1 \ldots x_k,)$, according to the generative model described above. It is the joint independent probability of the subsequences given the corresponding states and the probability of the transitions between states. That is,

$$P(X,\Phi) = P(q_1 \mid B)\left(\prod_{i=1}^{k} P(x_i \mid q_i)\right)\left(\prod_{i=1}^{k-1} P\left(q_{i+1} \mid node(q_i)\right)\right)$$

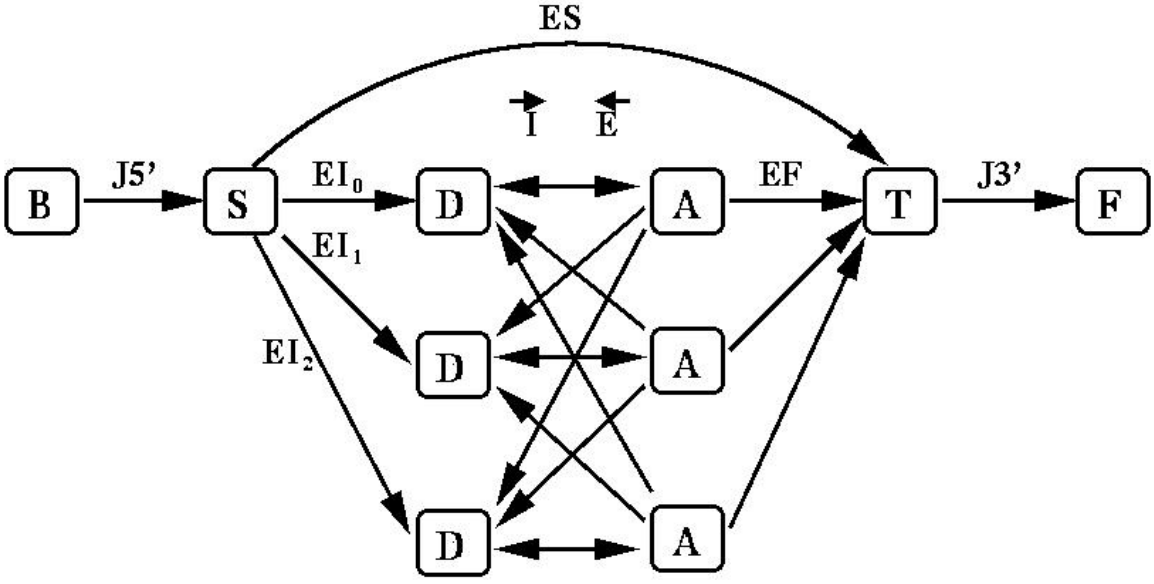
Given only the observed sequence *X,* using a variant of the Viterbi algorithm (Rabiner & Juang, 1986), we can calculate the parse Φ that maximizes the joint independent probability, i.e. the *most likely parse of X.* In a GHMM that represents a gene structure, such as the one in Figure 4-4, this most likely parse represents the model's prediction of the most likely gene structure within the sequence *X.* This variant of the Viterbi algorithm used to find the most likely parse is a dynamic programming algorithm that is essentially the same as the one defined by Auger and Lawrence (1989) to identify segment neighborhoods, by Sankoff (1992) optimally to decompose a sequence into disjoint regions with particular properties, and by Gelfand and Roytberg (1993), Snyder and Stormo (1993), Stormo and Haussler (1994), and many others to do gene finding, so I do not elaborate on it here. GHMMs place these previous approaches within a convenient and general probabilistic framework.

Figure 4-4: A simple GHMM for a sequence containing a gene. The arcs represent states that emit strings of bases and nodes represent transitions between states. The state labels are J5': 5' UTR, EI: Initial Exon, E: Exon, I: Intron, E: Internal Exon, EF: Final Exon, ES: Single Exon, and J3': 3' UTR. The node labels are B: Begin, S: Start Translation, D: Donor, A: Acceptor, T: Stop Translation, F: Final. The arrows imply a generation of bases from 5' to 3'.



The GHMM in Figure 4-4 represents only the basic ordering of gene features, and fails to capture fully the syntactic restrictions of a "legal gene parse." In an ideal DNA sequence, the parse is "frame consistent," i.e., the total number of coding nucleotides is a multiple of three and the reading frame is consistent from exon to exon. We can add additional states to the model graph such that only frame consistent parses are allowed. Figure 4-5 shows the model graph representing the resulting frame consistent GHMM. The three levels represent the three frames. Exon length can be restricted in the likelihood functions $P(x|q)$ to equal 0, 1 or 2 (modulo 3) for the various exon states in this GHMM in such a way to enforce frame consistency (for more details see (Kulp *et al*., 1997)). Another extension to the GHMM in the implementation, but not shown in the graph in Figure 4-5, is an arc leading back from node F to node B labeled with a state that generates intergenic non-coding bases between genes. In this way, the implemented structure allows multiple genes on both strands of the DNA. It has to be noted that this graph structure prohibits a more rare gene organization of genes within other genes.

Figure 4-5: A GHMM including frame constraints. "B" is the begin state, "J5'" the 5' UTR content sensor, "S" the start codon signal sensor, "EI" the initial exon content sensor, "D" the 5' splice site and "A" the 3' splice site sensor, "E" the internal exon content sensor, "I" the intron content sensor, "EF" the final exon content sensor, "T" the start codon signal sensor and "F" the end state. "ES" stands for the single exon gene content sensor. For multiple genes in genomic regions such as in the *Adh* region an additional arc loops from "F" to "B" and models the intergenic region including the promoter sensor.



4.3.1.1 Signal sensor models

Signal sensors are used to recognize transitions between states in a GHMM. This type of sensor is used in a pre-processing step to identify candidate sites where state transitions can occur. The dynamic programming method then uses the pre-processed transitions and finds the most likely parse. In the GHMMs shown in Figure 4-4 and Figure 4-5 the nodes correspond to gene features such as splice sites (GT and AG) and the positions of start and stop codons.

4.3.1.2 Content sensor models

Content sensors are used to estimate the likelihood of a subsequence given a particular state in the GHMM. Some basic content sensors used by Genie are described in detail in Kulp *et al.* (1996). Since that paper has been published, a more sophisticated type of content sensor has been developed. This type of content sensor integrates evidence from multiple sources and estimates a likelihood of a subsequence from the combined information.

In the Genie content sensor, each source of evidence is called a component; a component is trained to recognize a specific feature. Figure 4-6 shows an example of a fictitious subsequence whose likelihood is being evaluated by an internal exon content sensor. The internal exon content sensor is composed of several components: a nucleotide component, a codon component, end-region components representing the regions adjacent to the 5' and 3' splice sites, and a database homology match component. A component returns a likelihood for each potential feature occurrence, called an "extent." In the figure, the maximum likelihood is determined by the joint probability of the extents shown at the bottom of the figure, i.e. a 3' splice site extent, followed by two nucleotide extents, a database match extent, and three codon extents. Again, dynamic programming is used to decompose the subsequence into a series of extents in such a way that the joint probability of all extents is maximized. This decomposition is then used to calculate the likelihood. In addition to this likelihood a length distribution is added to each content sensor.

This simple, efficient method is a modular approach to developing an effective gene finding system because components can be easily added to or subtracted from a content sensor.

Figure 4-6: A sample content sensor (coding exon) combines evidence from multiple components to derive a maximum likelihood of the sequence. The arrow shows the combination of component features corresponding to the maximum likelihood.

4.4 Implementation of Genie

The Genie program has been developed over many years. The first implementation was from the original work first described in Kulp *et al.* (1996). This initial version was trained and optimized for human genes. Improvements on the splice site models as well as a description of the training for *Drosophila melanogaster* and initial results for this organism were reported in Reese *et al.* (1997). In early 1997 the system was extended to

integrate homology information into the statistical gene finding framework described in (Kulp et al., 1997).

In 1999 a faster version was developed. This new version is also a very modular system that allows for easy integration of new nodes, arcs and sensors. It also adopted the GFF format (formerly known as the Gene Feature Finding format; (Bruskiewich *et al*., 1999)) as an exchange format for pre-computed external sensors to be integrated into the GHMM framework. In addition the training was automated to allow easy retraining of the system for other organisms.

4.4.1 EST/cDNA sequence integration

In Genie 2, EST/cDNA alignments are used to predict intron splice pairs. This program is called *GenieEST*. Using BLASTN (Altschul & Gish, 1996), pairs of hits to the same subject sequence were extracted. When such pairs are approximately contiguous in the subject sequence and aligned near GT/AG splice boundaries then an intron was predicted. The content sensor models for splice sites and introns are modified such that the probability was artificially raised for these so-called EST introns, effectively constraining the system to ensure that the introns were correctly annotated according to the EST/cDNA evidence. This system GenieEST was tested and evaluated in the GASP assessment and results are listed below.

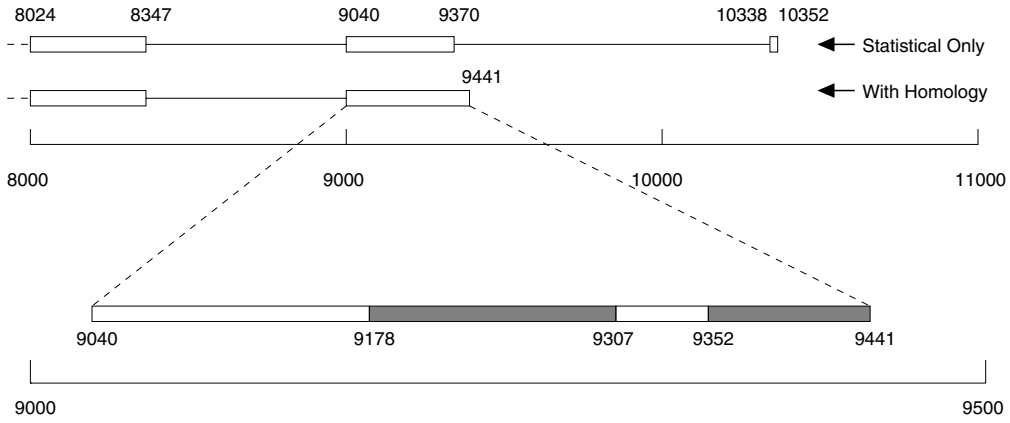
For the application to the complete genome of *Drosophila*, Genie's underlying graph model was extended to integrate information from 5' and 3' EST sequences from the same sequenced clone. The model was changed such that GenieEST was constrained so that it could not introduce intergenic regions between neighboring 5' and 3' EST alignments from the same cDNA clone.

4.4.2 Protein homology integration

Protein sequence homology is included as part of the content sensor for protein coding regions (described in detail in Kulp *et al.* (1997)). The corresponding program name as submitted to GASP is *GenieESTHOM*. Using BLASTX, candidate homologues are identified and assigned a likelihood probability similar to the Blast “S” score. The likelihood of a coding region that includes a protein database hit may be

higher than by statistical analysis alone depending on the degree of similarity. Figure 4-7 shows a typical refined prediction using homology information.

Figure 4-7: The diagram shows the gene prediction for the final 3,000 bases of the 11Kb DNA GenBank entry (Accession D14813). The first prediction is the result of running Genie without homology information. Genie fails to identify the complete final exon and predicts an additional small final exon. The second prediction includes BLASTX searches (GenieESTHOM) against the "nr" protein database. Here, two segments from a strong homologue (PIR Accession A38646) are found, shown shaded, with a small insertion between them. Using the additional information, the GenieESTHOM prediction is now correct.



4.4.3 Promoter neural network integration into Genie

Part of the gene structure GHMM submitted to GASP includes the core promoter region. The content sensor for this region is the time-delay neural network NNPP described in Chapter 2. The modularity of Genie was used to introduce this additional content sensor. For an unknown sequence the transcription start sites are pre-calculated and Genie then determines which core promoter fits best into its parse. This promoter site is then reported in the output file. The low specificity of independent promoter prediction (see Chapter 2) is compensated in this approach by integrating promoter prediction into the complete gene prediction. Thus, in effect, possible promoter sites are only considered upstream of a probable coding region.

Chapter 5 Results of Genie in *Drosophila*

This chapter addresses the general problem of evaluating computational annotation systems and gives results of the application of Genie in the *Adh* test bed sequence. In particular, Section 5.1 describes an annotation assessment project by reviewing several measures of the accuracy of gene and promoter prediction programs at the nucleotide, exon and gene levels. In addition to these traditional measures a novel statistic for measuring gene assembly tendencies is introduced. Problems in the final evaluation process are critically discussed. A discussion on the necessity of visualization tools for assessing and studying annotation methods concludes this section. Sections 5.2 covers the results of the application of Genie in *Adh.* Genie's performance is compared to that of other programs tested in the GASP experiment. The following Sections 5.3 - 5.4 take a closer look at the accuracy of the program addressing some of the strength and weaknesses in the current system. A collection of predicted novel genes in *Adh* is listed. Section 5.5 gives the results of the integrated promoter prediction in Genie and Section 5.6 addresses specific problems with the gene structure model applied in the GASP experiment. Resulting final improvements of the final program are presented.

5.1 Evaluating gene prediction

To assess genome annotations objectively, we need a test sequence that satisfies the following three requirements:

- The "correct answer", in our case the location of genes, has to be known.

- The underlying sequence region has to be representative of the entire genome.

- Meaningful evaluation statistics that describe the performance of programs have to be formulated.

In the GASP experiment we tried to do exactly that. Here I will give a short introduction to this experiment and discuss some of the evaluation statistics that were developed during the experiment.

The GASP experiment, the first of its kind, was similar in many ways to the CASP (Critical Assessment of Techniques for protein structure prediction) contests for protein structure prediction (Dunbrack *et al*., 1997; Levitt, 1997; Moult *et al*., 1997; Moult *et al*., 1999; Sippl *et al*., 1999; Zemla *et al*., 1999). However, unlike the CASP contest and following the famous fly tradition, GASP was promoted as a collaboration to evaluate various techniques for genome annotation.

Participants were given the finished genomic sequence for the *Adh* region and some related training data, but they did not have access to the full-length cDNA sequences that were sequenced for the paper by Ashburner *et al.* (1999) that describes the *Adh* region in depth. The experiment was widely announced and open to any participants. Submitters were allowed to use any available technologies and were encouraged to disclose their methods. Since a large group of participants was attracted and they provided a wide variety of annotations, GASP was able to assess the state of art in genome annotation.

Twelve groups participated in GASP, submitting annotations in one or more of six categories: *ab initio* gene finding, promoter recognition, EST/cDNA alignment, protein similarity, repetitive sequence identification and gene function.

5.1.1 Two standard annotation sets for the same *Adh* region (GASP)

We assembled two sets of gene annotations for the *Adh* region to use for evaluating gene prediction (see Appendix A for URL). The first standard annotation set, known as the *std1* data set, used high quality sequence from a set of 80 full-length cDNA clones from the *Adh* region to provide a standard with annotations that are very likely to be correct but certainly are not exhaustive. The second standard annotation set, known as the *std3* data set, was built from the annotations being developed for Ashburner *et al.* (1999) to give a standard with more complete coverage of the region, although with less confidence about the accuracy and independence of the annotations. *Std1* was a subset of *std3*.

The gene prediction evaluations focused on annotations that are specific to the coding region, from the start codon through the various intron-exon boundaries to the stop codon, and on promoter annotations.

The goal for our first standard set, *std1*, was to build a set of annotations that were believed very likely to be correct in their fine details (e.g. exact locations for splice sites), even if it did not include every gene in the region. S*td1* is based on alignments of 80 high quality, full-length cDNA sequences from this region with the high quality genomic sequence for the contig. The cDNA sequences are the product of a large cDNA sequencing project at the Berkeley *Drosophila* Genome Project and had not been submitted to GenBank at the time of the experiment. Working from five cDNA libraries, the longest clone for each unique transcript was selected and sequenced to a high quality level reaching an error rate of less than 1 in 10,000 bases. Starting with these cDNA sequences, alignments were generated to the genomic sequence using sim4 (Florea *et al*., 1998) and then filtered on several criteria (for details see the original GASP publication (Reese et al., 2000)). This process resulted in forty-three sequences from the *Adh* region for which structures were confirmed by alignments of high quality cDNA sequence data with an error rate again of less than 1 in 10,000 bases with high quality genomic data and by the fit of their splice sites to a *Drosophila* splice site model. Of these forty-three sequences, seven had a single coding exon and thirty-six had multiple coding exons. Start codon and stop codon annotations for these structures were added from the corresponding records in the *std3* data set.

The goal for the second standard set, called *std3*, was to build the most complete set of annotations possible while maintaining some confidence about their correctness. Ashburner *et al.* (1999) compiled an exhaustive and carefully curated set of annotations for this region of the *Drosophila* genome based on information from a number of sources, including BLASTN, BLASTP (Altschul et al., 1990), and PFAM alignments (Bateman et al., 2000; Sonnhammer *et al*., 1998; Sonnhammer *et al*., 1997), high scoring GENSCAN (Burge & Karlin, 1997) and Genefinder (Green, 1995) predictions, ORFFinder results (Friese *et al*., 1999), full length cDNA clone alignments (including those used in *std1*), and alignments with full length genes from GenBank. This set included 222 gene structures: 39 with a single coding exon, and 183 with multiple coding exons. Of these 222 gene structures, 182 are similar to a homologous protein in another organism or have a *Drosophila* EST hit. For these structures, the intron-exon boundaries were

verified by partial cDNA/EST alignments using sim4 (Florea et al., 1998), homologies were discovered using BLASTX, TBLASTX and PFAM alignments, and gene structure was verified using a version of GENSCAN trained for finding human genes. Of the fifty-four remaining genes, fourteen had EST or homology evidence but were not predicted by GENSCAN or Genefinder, and forty were based entirely on strong GENSCAN and Genefinder predictions. All of this evidence was evaluated and edited by experienced *Drosophila* biologists, resulting in a data set that exhaustively covers the region with a high degree of confidence.

Building a set for the evaluation of transcription start site or, more generally, for promoter recognition, proved to be even more difficult. For the genes in the *Adh* region almost no experimentally confirmed annotation for the transcription start site existed. As the 5' UTR regions in *Drosophila* can extend up to several kilobases, the region directly upstream of the start codon could not be used. To obtain the best possible approximation, the 5' ends of annotations from Ashburner *et al.* (1999) were taken, where the upstream region relied on experimental evidence (the 5' ends of full-length cDNAs) and for which the alignment of the cDNA to the genomic sequence included a good open reading frame. The resulting set contained 92 genes out of the 222 annotations in the *std3* set (Ashburner et al., 1999). This number is larger than the number of cDNAs used for the construction of the *std1* set described above because cDNAs that were already publicly available were included. The 5' UTR of these 96 genes has an average length of 1,860 basepairs, a minimum length of 0 basepairs (when the start codon was annotated at the beginning) and a maximum length of 36,392 basepairs.

5.1.2 Evaluation statistics for gene finding

An ideal gene prediction tool would produce annotations that were exactly correct and entirely complete. The fact that no existing tool has these characteristics reflects our incomplete understanding of the underlying biology as well as the difficulty of building adequate gene models in a computer. While no tool is perfect, each tool has particular strengths and weaknesses and any performance evaluation should be in the context of an intended use. In one of the first reviews of gene prediction accuracy, Fickett and Tung (1992) developed a method that measured predictors' ability to correctly recognize coding regions in genomic sequence. They used their method to compare published techniques and concluded that in-frame hexamer counts were the

most accurate measure of a region's coding potential. Burset and Guigó (1996) recognized that there are a wide variety of uses for gene predictions and developed measures--including base level, exon level, and gene level specificity and sensitivity--that describe a predictor's suitability for a particular task.

When assessing the accuracy of predictions, each prediction falls into one of four categories. A true positive (TP) prediction is one that correctly predicts the presence of a feature (Figure 5-1). A false positive (FP) prediction incorrectly predicts the presence of a feature (Figure 5-2). A true negative (TN) prediction is correct in not predicting the presence of a feature when it isn't there. A false negative (FN) prediction fails to predict the existence of a feature that actually exists (Figure 5-3). The sensitivity of a tool is defined as TP/(TP+FN), and can be thought of as a measure of how successful the tool is at finding things that are really there. The specificity of a tool is defined as TP/(TP+FP), and can be thought of as a measure of how careful a tool is about not predicting things that aren't really there. Burset and Guigó (1996) also use a correlation coefficient and an average correlation coefficient. These measures were not used in GASP because they depend on predictors' false negative information and it is recognized that the evaluation sets were constructed in such a way that the false negative information is not trustworthy.

Figure 5-1: True positives (TP). True positive predictions at the gene level against the *std1* data set (shown right below the *Adh* sequence bar) are demonstrated for two fictional prediction programs, "pred 1" and "pred 2". "Pred 1" predicts 3 genes and "pred 2" predicts seven genes. In *std1* three genes are annotated. The true positive genes for "pred 1" are noted right below the shadowed boxes crossing all three predictions with the "TP" label. "Pred 1" has two true positive predictions (genes 2 and 3 in *std1*). "Pred 2" has also two true positive predictions ("TP" labels on the bottom of the figure) but these are genes 1 and 2 in the *std1* annotation set.

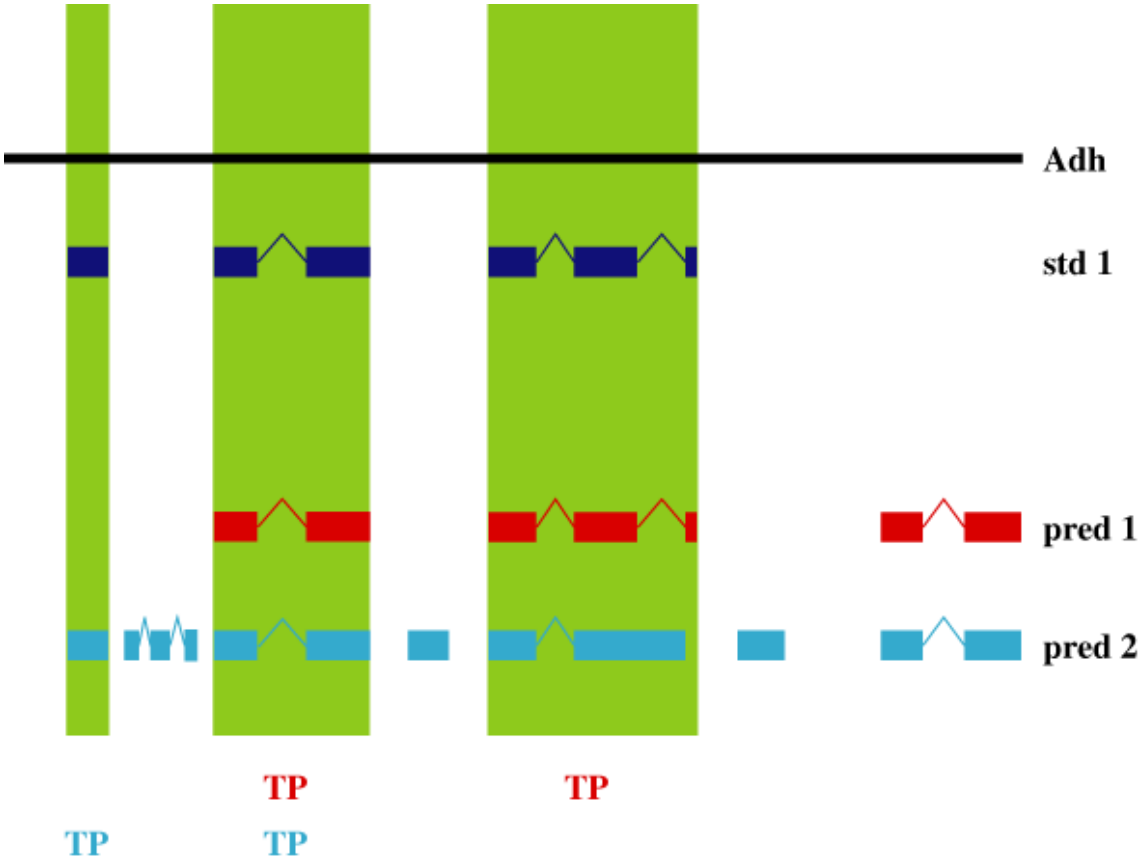


Figure 5-2: False positives (FP). The false positive predictions at the gene level against the *std1* data set are demonstrated. "Pred 1" has one false positive predicted gene and "pred 2" has five.

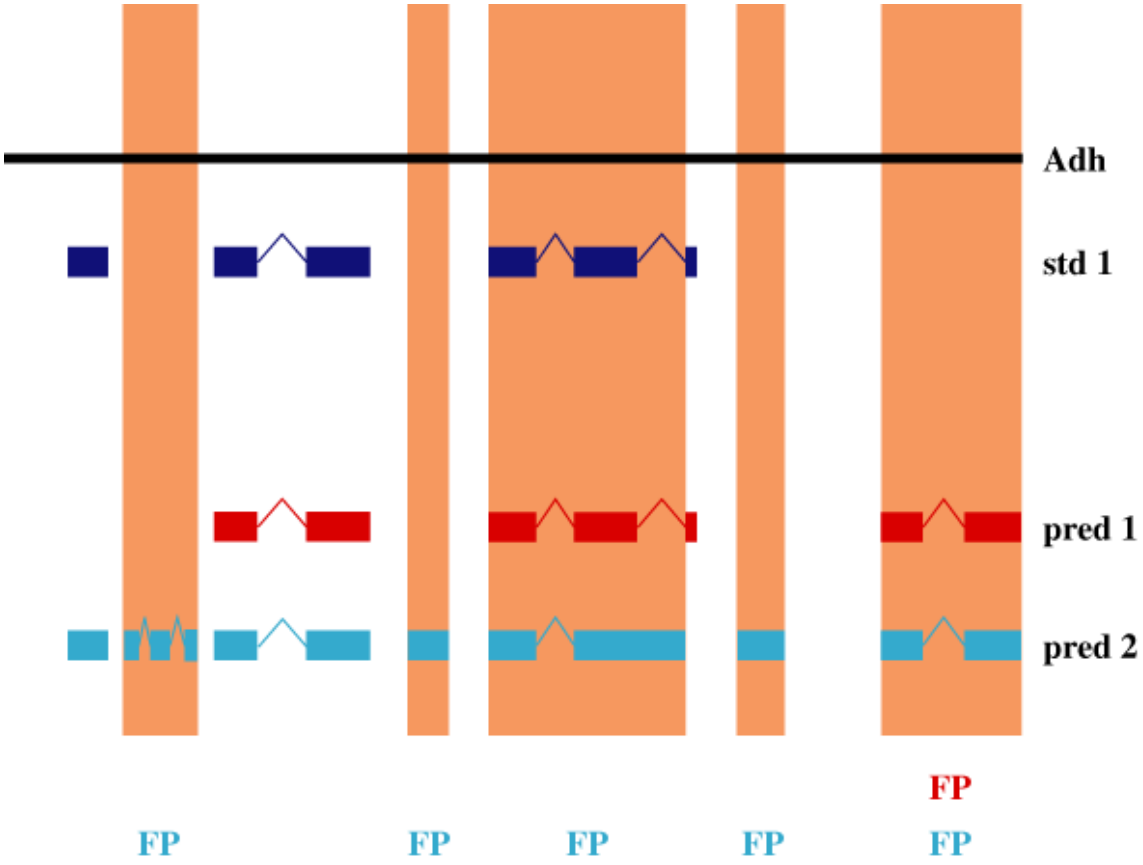
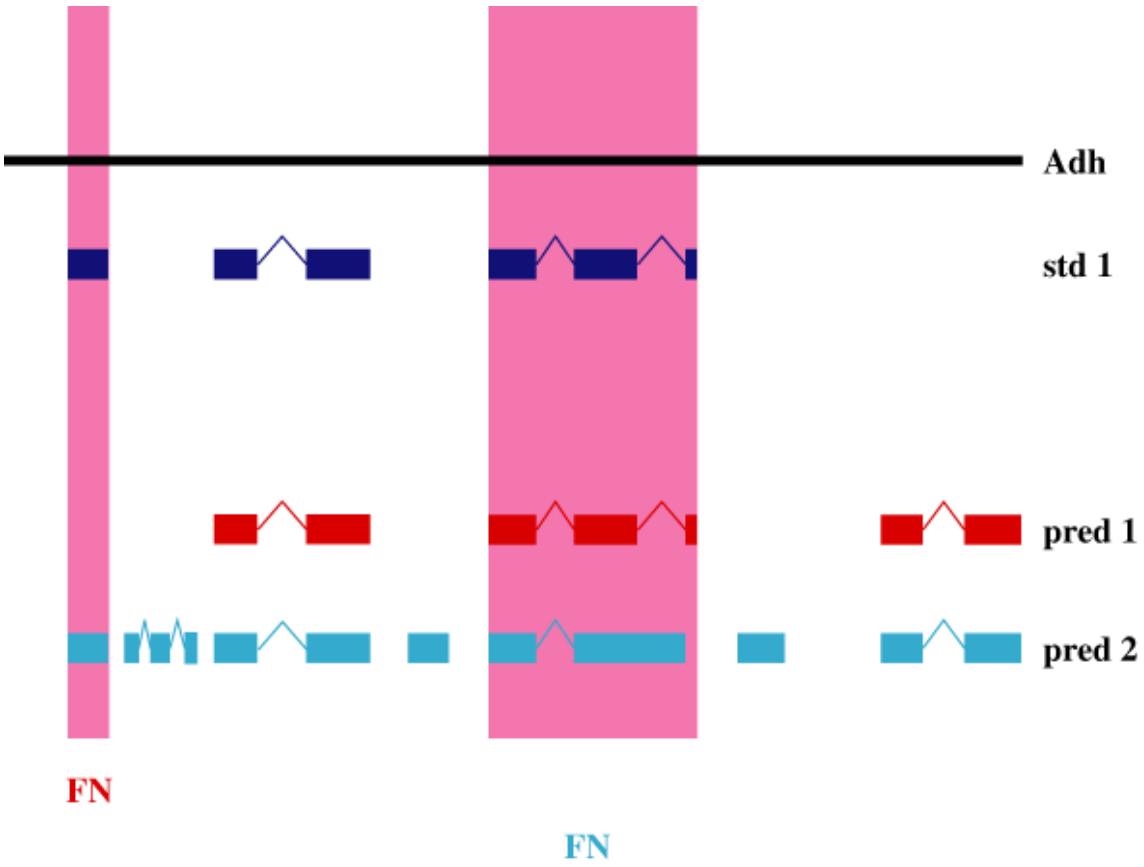


Figure 5-3: False negative (FN). The false negative predictions at the gene level against the *std1* data set are demonstrated. Both "pred 1" and "pred 2" have one false negative predicted gene.



5.1.3 Base level

The base level score measures whether a predictor is able to correctly label a base in the genomic sequence as being part of some gene. It rewards predictors that get the broad sweeps of a gene correct, even if they don't get the details such as the splice site boundaries entirely correct. It penalizes predictors that miss a significant portion of the coding sequence, even if they get the details correct for the genes they do predict. Sensitivity and specificity measures were used as defined above as the measures of success in this category.

5.1.4 Exon level

Exon level scores measure whether a predictor is able to identify exons and correctly recognize their boundaries. Being off by a single base at either end of the

exon makes the prediction incorrect. Since only coding exons are considered, the first exon is bracketed by the start codon and a 5' splice site, the last exon is bracketed by a 3' splice site and the stop codon, and the interior exons are bracketed by a pair of splice sites. As measures of success in this category, two statistics in addition to sensitivity and specificity are used. The missed exon score is a measure of how frequently a predictor completely failed to identify an exon (no prediction overlap at all), while the wrong exon score is a measure of how frequently a predictor identifies an exon that has no overlap with any exon in the standard sets. The missed exon score is the percentage of exons in the standard set for which there were no overlapping exons in the predicted set. Similarly, the wrong exon score is the percentage of exons in the predicted set for which there were no overlapping exons in the standard set.

5.1.5 Gene level

Gene level sensitivity and specificity measure whether a predictor is able to correctly identify and assemble all of a gene's exons. For a prediction to be counted as a true positive, all of the coding exons must be identified, every intron-exon boundary must be exactly correct, and all of the exons must be included in the proper gene. This is a very strict measure that addresses a tool's ability to perfectly identify a gene. In addition to the sensitivity and specificity measures based on absolute accuracy, the *missed genes* score is used as a measure of how frequently a predictor completely missed a gene (a gene is considered missed if none of its exons are overlapped by a predicted coding gene; see Figure 5-4) and the *wrong genes* score is used as a measure of how frequently a predictor incorrectly identified a gene (a prediction is considered wrong if none of its exons are overlapped by a gene from the standard set; see Figure 5-5).

Figure 5-4: Missed genes (MG). "Pred 1" does not overlap the first gene in *std1*. Therefore it is marked as a missed gene. "Pred 2" does not miss any gene from *std1*.

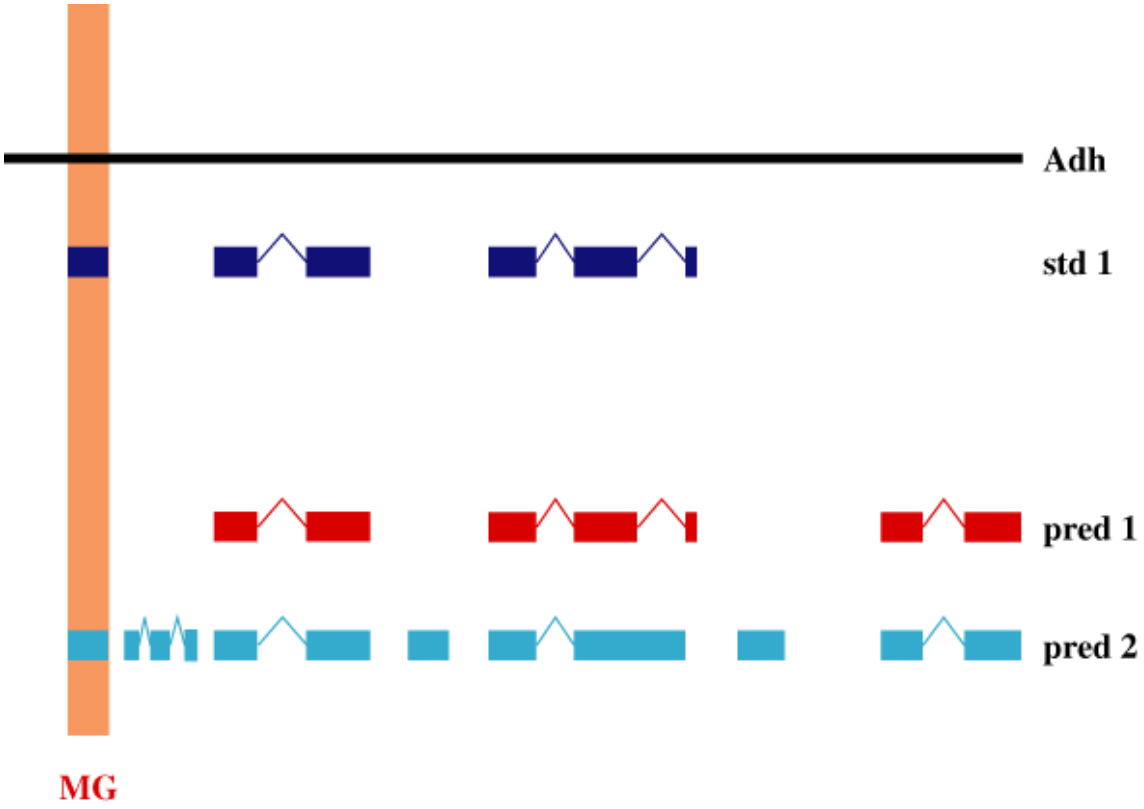
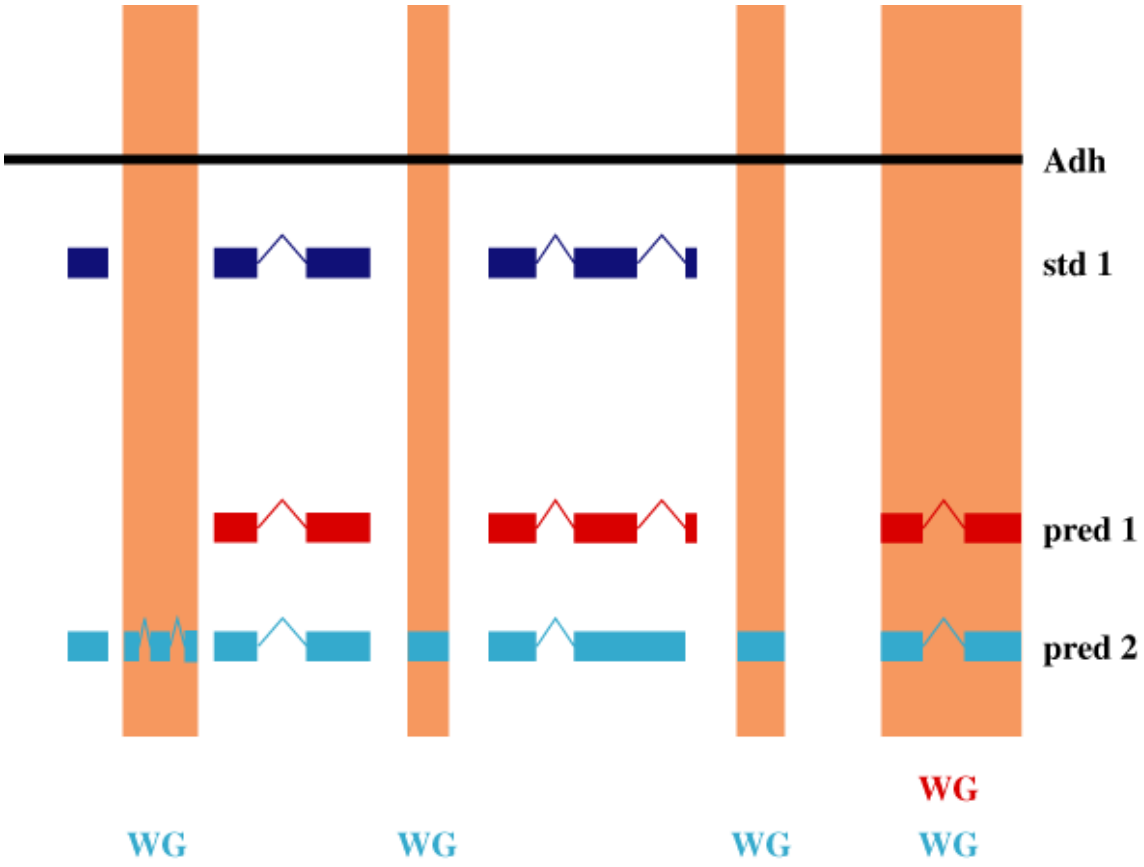


Figure 5-5: Wrong genes (WG). "Pred 1" and "pred 2" both have an extra predicted gene at the end of the example sequence that does not overlap any gene in *std1*. Therefore these predictions are counted as wrong genes. In addition, "pred 2" has three more wrong genes.



5.1.6 Split and Joined genes

The exon level scores discussed above measure how well a predictor recognizes exons and gets their boundaries exactly correct. The gene level scores measure how well a predictor can recognize exons and assemble them into complete genes. Neither of these scores directly measures a predictor's tendency to incorrectly assemble a set of predicted exons into more or fewer genes than it should. During GASP we developed two new measures, *split genes* and *joined genes,* which describe how frequently a predictor incorrectly splits a gene's exons into multiple genes and how frequently a predictor incorrectly assembles multiple genes' exons into a single gene. Because the coverage of the *std1* data set is so incomplete, split genes and joined gene scores are only included from the comparison with *std3*. A gene from the standard set is considered *split* if it overlaps more than one predicted gene (see Figure 5-6). Similarly, a predicted gene is considered *joined* if it overlaps more than one gene in the standard

set (see). The split genes measure is defined as the sum of the number of predicted genes that overlap each standard gene divided by the number of standard genes that were split. The joined genes measure is the sum of the number of standard genes that overlap each predicted gene divided by the number of predicted genes that were joined. A score of 1 is perfect and means that each of the genes from one set overlaps exactly one gene from the other set.

Figure 5-6: Split genes (SG). The toy example shown in Figure 5-1 to Figure 5-5 is modified to demonstrate the measures for split and joined genes. Here "pred 2" breaks up the annotated second gene in *std1* into two separate single exon genes (the splitting is demonstrated with the shaded box). Therefore this prediction is counted as a split gene.

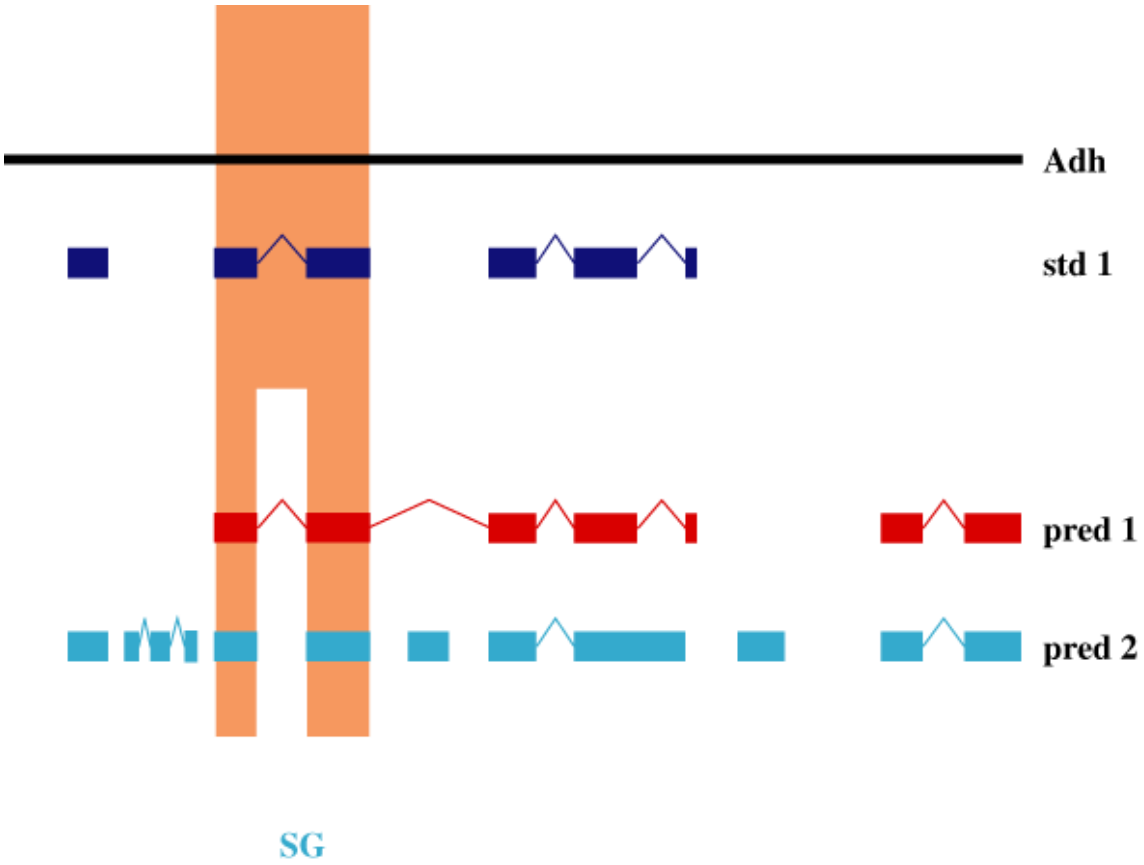
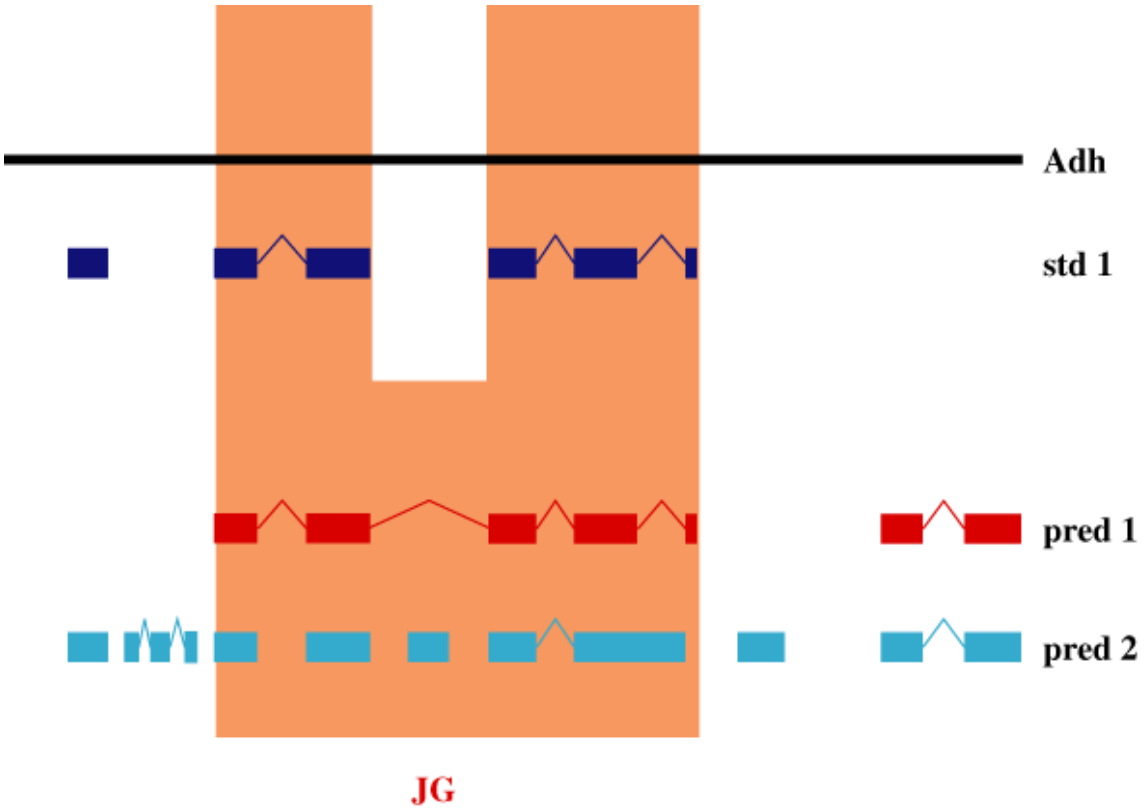


Figure 5-7: Joined genes (JG). "Pred 1" merges the annotated gene 2 and 3 in the *std1*set into one long gene. For "pred 1" this is counted as a joined gene (the join is demonstrated with the shaded box).



5.1.7 Application of these measures to "correct answer" data sets *std1/std3*

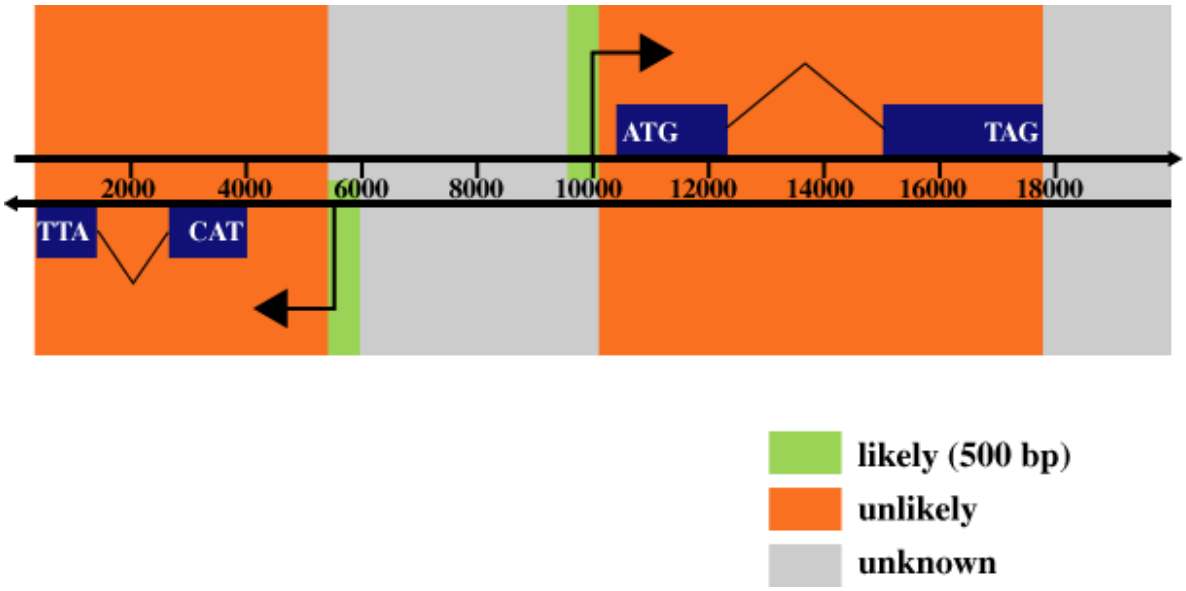
The *std1* dataset is built in such a way that it should be correct in the details of the genes that it describes, though it is clear that it only includes a small portion of the genes in the region. The *std3* data set, on the other hand, is as complete as was possible, but does not have rigorous independent evidence for all of its annotations. For the *std1* dataset, the TP count (it was predicted and it exists in the standard) and FN count (it was not predicted but it does exist in the standard) are reliable because of the confidence in the correctness of the predictions in the set. On the other hand, the TN count (it was not predicted and it is not in the standard set) and FP count (it was predicted but is not in the standard set) are not reliable because they both assume that the standard correctly describes the absence of a feature and it is known that there are genes missing from *std1*. It follows that sensitivity is meaningful for *std1* because it only depends on TP and FN but that one can be less confident about the specificity score, since it depends on TP and FP. A similar logic applies to the *std3* dataset, where the confidence in the set's

completeness but not its fine details suggests that the TP and FP scores are usable but that the TN and FN scores are not. This means that for *std3*, the specificity measure can be used to describe a predictor's performance but that sensitivity is likely to be misleading.

5.1.8 Evaluation of promoter predictions

For evaluating promoter predictions the measures proposed by Fickett and Hatzigeorgiou (1997) were adopted. They evaluated the success of promoter predictions by calculating the percentage of correctly identified transcription start sites versus the false positive rate. A TSS is regarded as correctly identified if a program makes one or more predictions within a certain "likely" region around the annotated site. The false positive rate is defined as the number of predictions within the "unlikely" regions outside the "likely" regions divided by the total number of bases contained in the unlikely set. As the annotation of the TSS is only preliminary and not experimentally confirmed, a rather large region of 500 bases upstream and 50 bases downstream of the annotated TSS is chosen as the "likely" region (see Figure 5-8). The upstream region is always taken as the "likely" region, even if it overlaps with a neighboring gene annotation on the same strand. The "unlikely" region for each gene thus consists of the rest of the gene annotation, from base 51 downstream of the TSS to the end of the final exon.

Figure 5-8: Promoter prediction in *Adh*. Two toy genes are shown in the schematic, one on the forward and one on the reverse strand. The shadowing demonstrates the classification of the sequence into "likely" regions (+500 to -50 around the TSS; medium shadowing), "unlikely" region, overlapping the coding gene (dark shadowing) and "unknown" region (light shadowing) for the rest of the sequence.



5.1.9 Visualization of the annotations

Generating "good" annotations generally requires integrating multiple sources of information, such as the results of various sequence analysis tools plus supporting biological information. Visualization tools that display sequence annotations in a browsable graphical framework make this process much more efficient. Visualization tools are essential in order to evaluate genome annotations. When annotations are displayed visually, overall trends become apparent, for example gene-rich vs. gene-poor regions. During all my work I was fortunate in that the Berkeley Drosophila Genome Project had built a flexible suite of genome visualization tools (Helt & al., 1999) that were used to build a genome annotation browser called CloneCurator (Harris *et al.*, 1999).

CloneCurator (Figure 5-9) displays features on a sequence as colored rectangles. Features on the forward strand appear above the axis, while those on the reverse strand appear below the axis. The display can be zoomed and scrolled to view areas of interest in more detail. A configuration file identifies the feature types that are to be displayed,

and assigns colors and offsets to each one. CloneCurator was used to assess the GASP predictions.

5.2 Accuracy of Genie in *Adh*

In the GASP experiment on the *Adh* sequence three sets of predictions generated by Genie were assessed. The first, named *Genie*, was generated using the GHMM trained on the cleaned gene collection described in 4.2 based on statistical information only. The second set, named *GenieEST*, was generated using the same signal sensors as *Genie* but extended the content sensors by incorporating EST information for the determination of the splice boundaries. The third prediction, named *GenieESTHOM*, used, in addition to all the models from *GenieEST*, protein homology information from BLAST runs (Altschul & Gish, 1996) against the non-redundant protein Genbank database (nr). This run resulted in DNA-protein alignments to related protein sequences in *Drosophila melanogaster*, as well as to related protein sequences in other organisms.

Table 5-1: GASP Evaluation of gene finding systems (taken from (Reese et al., 2000)). The evaluation is divided into three categories: Base level, exon level and Gene level. The statistics reported are Sensitivity (Sn), Specificity (Sp), Missed Exons (ME), Wrong Exons (WE), Missed Genes (MG), Wrong Genes (WG), Split Genes (SG) and Joined Genes (JG). "*Std1*" and "*std3*" indicate against which standard set the statistics are reported.

		Fgenes CGG1	Fgenes CGG2	Fgenes CGG3	Gene ID v1	Gene ID v2	Genie	Genie EST	Genie EST HOM	HMM Gene	MAGPIE exon	Grail exp
Base level	Sn *std1*	0.89	0.49	0.93	0.48	0.86	0.96	0.97	0.97	0.97	0.96	0.81
	Sp *std3*	0.77	0.86	0.60	0.84	0.83	0.92	0.91	0.83	0.91	0.63	0.86
Exon level	Sn *std1*	0.65	0.44	0.75	0.27	0.58	0.70	0.77	0.79	0.68	0.63	0.42
	Sp *std3*	0.49	0.68	0.24	0.29	0.34	0.57	0.55	0.52	0.53	0.41	0.41
	ME(%) *std1*	10.5	45.5	5.6	54.4	21.1	8.1	4.8	3.2	4.8	12.1	24.3
	WE(%) *std3*	31.6	17.2	53.3	47.9	47.4	17.4	20.1	22.8	20.2	50.2	28.7
Gene level	Sn *std1*	0.30	0.09	0.37	0.02	0.26	0.40	0.44	0.44	0.35	0.33	0.14
	Sp *std3*	0.27	0.18	0.10	0.05	0.10	0.29	0.28	0.26	0.30	0.21	0.12
	MG(%) *std1*	9.3	34.8	9.3	44.1	13.9	4.6	4.6	4.6	6.9	4.6	16.2
	WG(%) *std3*	24.3	24.8	52.3	22.2	30.5	10.7	13.0	15.5	14.9	55.0	23.7
	SG	1.10	1.10	2.11	1.06	1.06	1.17	1.15	1.16	1.04	1.22	1.23
	JG	1.06	1.09	1.08	1.62	1.11	1.08	1.09	1.09	1.12	1.06	1.08

In the *Adh* region, Genie, GenieEST and GenieESTHOM predicted a total of 241, 246 and 258 genes, respectively. In general all three programs scored well in the gene finding category (Table 5-1, taken from Reese *et al.* (2000)). The summary of the results is divided into the three categories of Base level, Exon level and Gene level and the performance discussed for all three versions of Genie.

5.2.1 Base level

The statistical Genie program achieves 96% sensitivity (Table 5-1). The extra information from ESTs and homology improves the sensitivity of the statistical Genie outcome by 1% to 97%. Most of the bases belonging to coding exons seem to be

predicted by Genie, which makes the tool robust and sensitive for a first scan of genomes to identify most of the proteome of an organism.

In specificity one can see a drop in performance for the Genie annotations that use homology information (*GenieESTHOM*) to 83% from 92% for *Genie* and 91% for *GenieEST* respectively. This is surprising and means that GenieESTHOM uses misleading protein homology information to incorrectly predict coding regions that are non-coding. This is due to some weaker homology hits that indicate similarities to protein-like elements in the DNA. These hits could be due to pseudo-genes or just simply to elements that are protein like and were originally derived from real protein sequences either through outside integration by transposons, viruses or by evolutionary gene duplication and subsequent degeneration through mutations. For thirteen of the over-predicted genes it is clear that they overlap transposable elements and therefore all thirteen are counted as false positives (see Table 6 for details on the overlapped transposons).

5.2.2 Exon level

Predicting splice sites, translation initiation and termination is difficult to accomplish within a purely computational framework because these sites can be very divergent and might be regulated through the over- or under-representation of nucleotides in the respective consensus sequences. Prediction is further confounded by external enhancer or repressor binding sites that are not well understood. The low rate of *missed exons* of 8.1%, 4.8% and 3.2% for *Genie*, *GenieEST* and *GenieESTHOM,* respectively, and the high sensitivity scores of over 70% suggest that Genie finds almost all the exons but has more trouble predicting the precise boundaries correctly. *GenieEST* demonstrates significant improvement (sensitivity of 77% compared to 70%) in splice site identification, which is to be expected from the EST alignments. Sensitivity improves to 79% in *GenieESTHOM*. This tendency of improved scores for *GenieEST* and *GenieESTHOM* reverses itself on the specificity scores and *wrong exon* scores. Here the best scores are from the pure statistical Genie program. This might reflect the source of data in the *std3* reference set of presumed correct gene structures, where quite a number of the genes are based on pure GENSCAN (Burge & Karlin, 1997) predictions, a program similar in structure and concept to the statistical Genie program.

5.2.3 Gene level

All three versions of Genie have problems assembling complete genes absolutely correctly. It is clear that this is a very hard problem, and so we find a sensitivity of 44% by *GenieEST* and *GenieESTHOM* to be very promising. This is due to a well-balanced integration of statistical sensors combined with the strength of sequence similarity methods. Specificity is almost equal for *Genie* and *GenieEST* but drops for *GenieESTHOM* due to misleading hits to low scoring protein-like elements. The relatively low number of *wrong genes* (10.7%) for the pure statistical *Genie* implies that users can have confidence that predicted genes do correspond, at least in part, to true protein coding regions. Nothing in the training of Genie or in the application constrained Genie from predicting the transposases and the reverse transcriptases in the transposable elements as genes. And, of course, there might also be new genes that Genie recognizes that are not yet in the biological annotation from *std3* (see Table 5-2 for details).

The statistic of *split genes and joined genes* describes the tendency of a program to assemble or split apart genes (see 5.1.6). The *split gene* numbers range from 1.17 to 1.15 for the three Genie programs, which indicates a high number of genes that are incorrectly split into several separate gene predictions. 15-17% of all genes are split into one or more separate gene predictions. The *joined gene* numbers are much lower (1.08-1.09) indicating the tendency of Genie to prefer to break up genes instead of joining them. Compared to other gene finders both numbers are high, suggesting that other programs have better solutions for this problem.

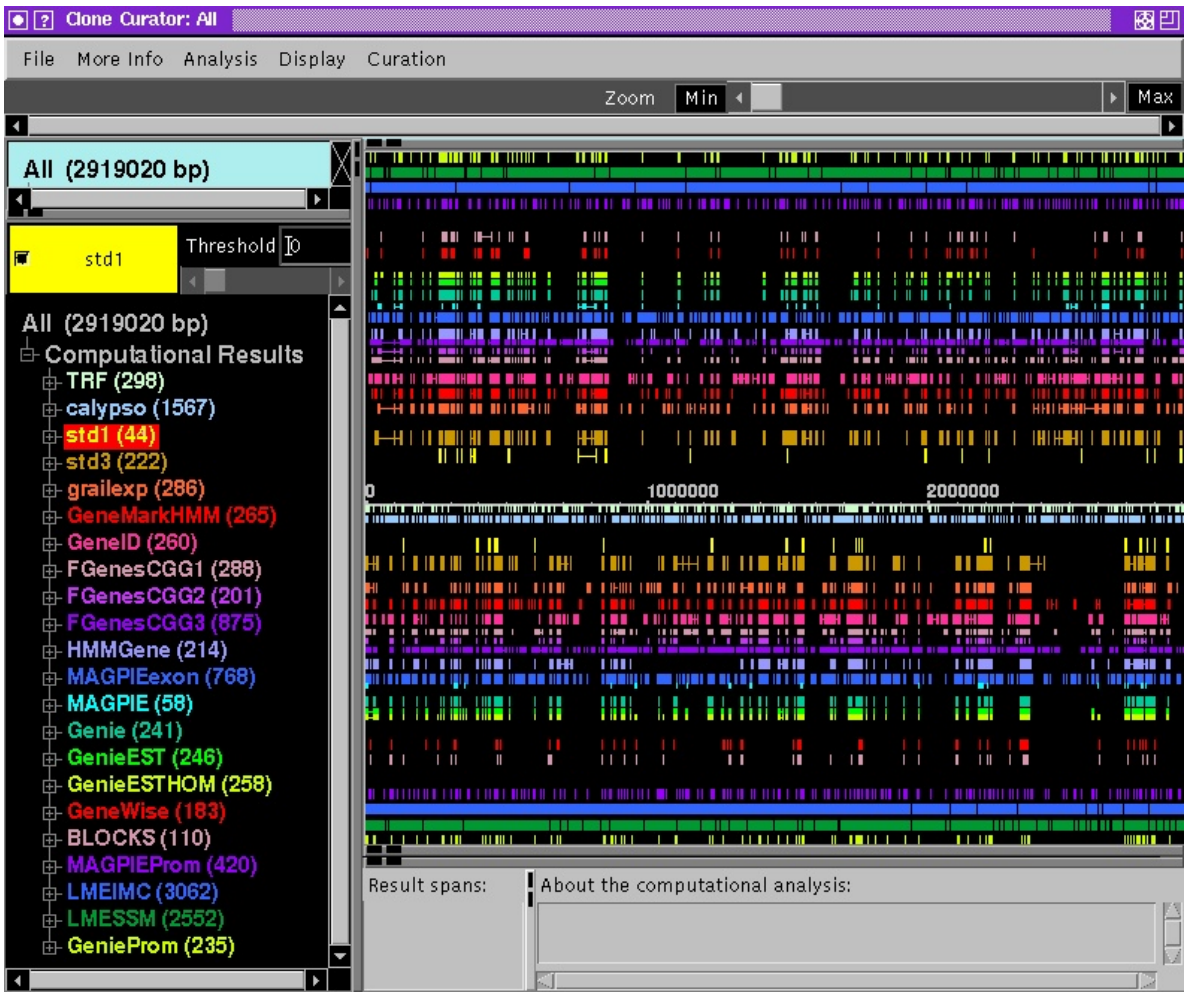
5.3 Selected Genie annotations in *Adh*

In this section we discuss selected predictions or missed predictions from *Genie*, *GenieEST*, *GenieESTHOM* and *GenieESTPROM* compared to the standard sets *std1* and *std3* as well as the behavior of Genie compared to other gene finding systems based on the selected examples from the GASP experiment (Reese et al., 2000).

Figure 5-9 shows the *Adh* region with all the submitted annotations to the GASP experiment, displayed by the program CloneCurator (Harris et al., 1999). As indicated in this figure and apparent in all subsequent genomic map figures, the three Genie

submissions are grouped together. They are the group of gene finders that are further from the genomic sequence axis, next to the protein homology annotations.

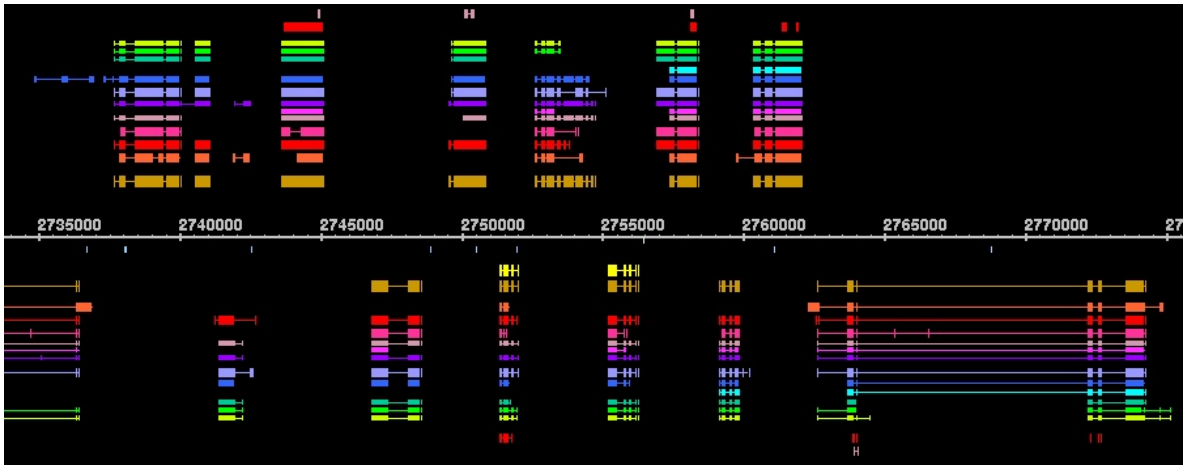
Figure 5-9: The *Adh* region. A screen shot of the annotated *Adh* region displayed by the CloneCurator program (Harris et al., 1999)**. It shows the genome annotations of all 12 groups in GASP. The main panel shows the computational annotations on the forward strand (above axis) and the reverse strand (below axis). Genes located on the top half of each map are transcribed from distal to proximal (with respect to the telomere of chromosome arm 2L); those on the bottom are transcribed from proximal to distal. Right below the axis are the two repeat finding results, followed by reference sets from Ashburner *et al.* (*std1* and *std3*), followed by the twelve submissions of gene finding programs, followed by the two protein homology programs and eventually, farthest away from the axis, the four promoter recognition programs. The left panel gives the color-coded legend for the programs and the number of predictions made by the programs.**



In the "busy" region, Figure 5-10, all three *Genie* programs predict the first four (*DS02740.4*, *DS02740.5*, *I(2)35Fb*, *DS02740.8*) and the last of the forward strand genes (*fzy*) correctly. The fifth gene (*DS02740.10*), between 2,752,000 and 2,755,000 is only

predicted by *GenieEST* and *GenieESTHOM*. This indicates that coding potential is not strong enough to distinguish protein coding from intergenic sequence and the additional information from EST alignments is necessary to identify the coding regions for this gene. Although *GenieEST* and *GenieESTHOM* predict the first three exons correctly, both miss the 3' splice site for the fourth exon and then select a 3' splice site in a different frame so that a stop codon is introduced in the middle of the real fourth exon. Thus, both programs miss the last four exons. The coding potentials for these remaining four exons are low, which is suggested by the fact that nothing is predicted with the statistical *Genie*. The fifth gene (*Sed5*) in this region on the forward strand is very interesting. While two of the seven gene finding programs agree with the suggested annotation in *std3*, four others predict a longer first coding exon. All three Genie programs also predict a longer initial coding exon. This is very interesting due to the difficulty determining the exact start of translation of a gene. Most biologists assign the first "ATG" in a 5' EST sequence followed by a long open reading frame (ORF) as the real start codon, but this is not a strict rule and might be wrong in some cases.

Figure 5-10: "Busy" region. Annotations for the following known genes described in Ashburner *et al.* (1999) are shown for the region from 2,735,000 - 2,775,000 (from the left to the right of the map):
***crp* (partial, rev.), *DS02740.4* (f), *DS02740.5* (f), *I(2)35Fb* (f), *heix* (r), *DS02740.8* (f), *DS02740.9* (r), *DS02740.10* (f), *anon-35Fa* (r), *Sed5* (f), *cni* (r), *fzy* (f), *cact* (r).**

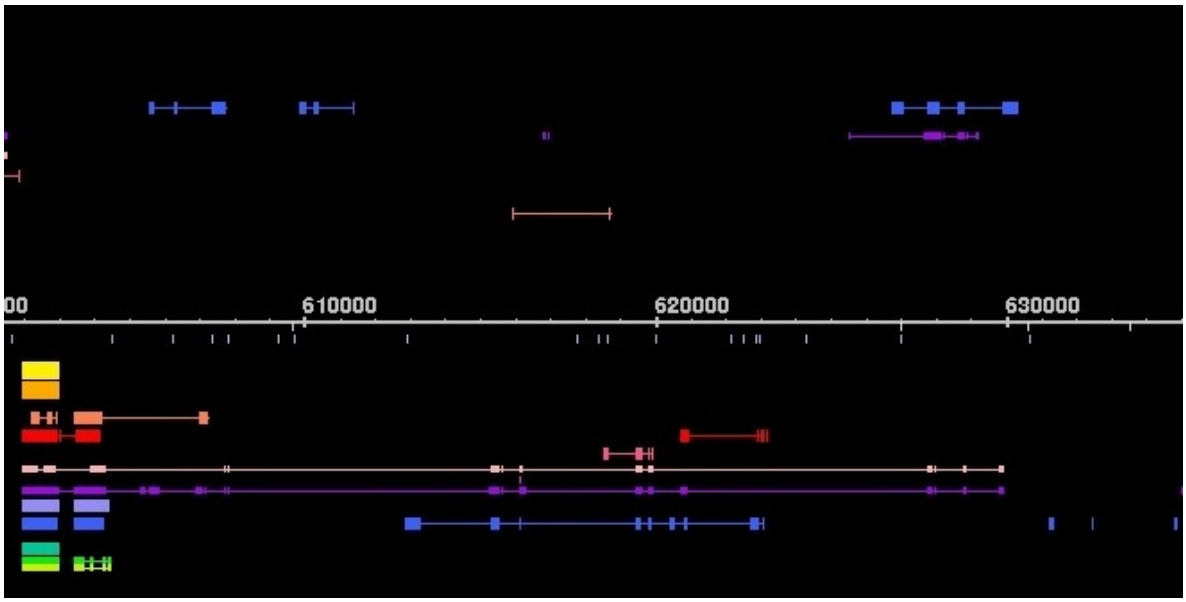


On the reverse strand the complete Genie suite predicts a two-exon gene at 2,741,000 - 2,742,500, where no gene exists in the *std3* reference set. Because this prediction agrees with four other gene finding programs and does not overlap any of the transposon annotations from Ashburner *et al.* (1999) this might be a real gene missed in

the *std3* set. This gene also does not show any protein homology and might therefore be a novel gene (see Table 1 for details). Further EST screening and subsequent full-length sequencing studies may confirm this hypothesis. All three Genie programs predict the next gene (*heix*) correctly, but the third gene (*DS02740.9*) on the reverse strand is not predicted. The statistical *Genie* misses the first two exons and introduces a wrong start codon. EST sequence information extends the *GenieEST* and *GenieESTHOM* predictions, correctly identifying the final two and the second exon. But both programs miss the initial exon, which is only three (!) basepairs long; the Genie model has a minimum length requirement of six bases. The fourth (*anon-35Fa*) and fifth (*cni*) genes on the reverse strand, both short genes with four and five exons respectively, are both predicted completely correctly. The last and longest gene (*cact* or *cactus*) in this region spans almost 12Kb from 2,762,639 - 2,774,287. The interesting fact about this gene is that it has a very long intron between the third and fourth exon spanning 8Kb. While most of the other gene finding programs predict this intron correctly, all three Genie programs miss this intron and split this gene into two separate genes. This is a typical behavior for Genie and is addressed in the next version of the program.

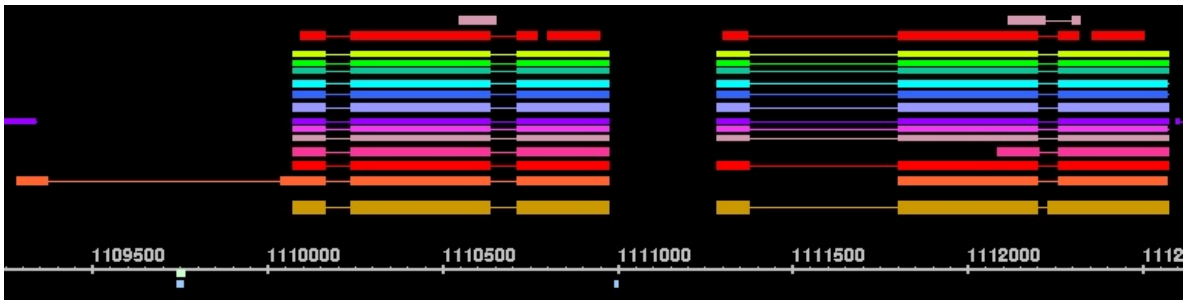
Genie's overall low false positive rate is demonstrated in the gene poor region, shown in Figure 5-11. In this "gene desert" Genie predicts only two genes both on the reverse strand. The first is a single exon gene (*DS01759.1*), which is correctly predicted by all three Genie programs. *GenieEST* and *GenieESTHOM* both agree on an additional gene after this first single exon gene. While other gene finding programs predict single exon genes or genes containing two exons here, *GenieEST* and *GenieESTHOM* predict a gene with four exons. While the exact structure of a possible gene in this region can only be wildly speculated, it seems probable that there is a novel gene in this region.

Figure 5-11: Gene desert. Annotations for the following known gene described in Ashburner *et al.* are shown for the region from 600,000 - 635,000 (from the left to the right of the map):
***DS01759.1* (r).**



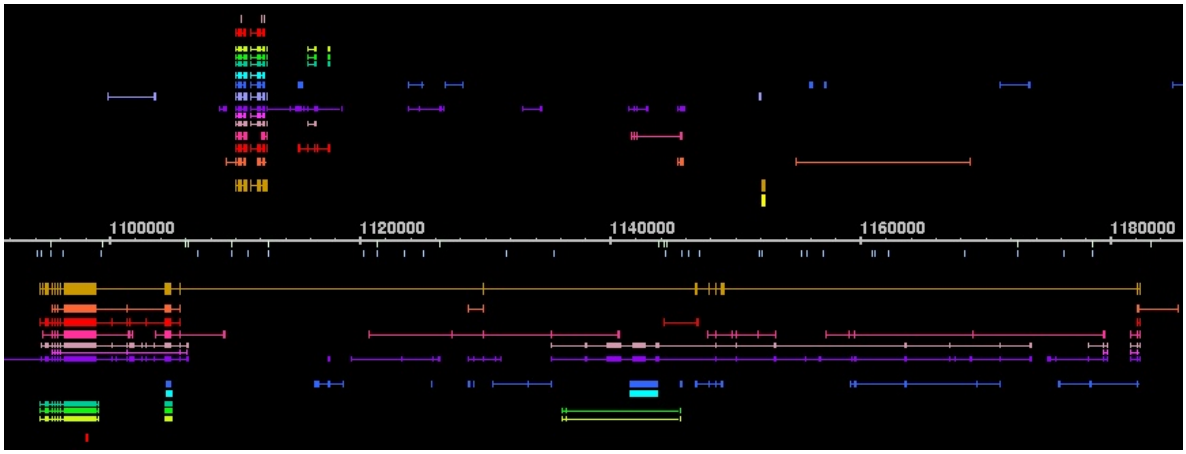
All Genie programs predict the genes *Adh* and *Adhr* (Figure 5-12), correctly. This is not surprising for *GenieEST* and *GenieESTHOM* because there are many ESTs available for both genes. But even without EST evidence, *Genie* predicted these duplicated genes correctly. As described in Ashburner *et al.* both genes are active but are both regulated by the same promoter. The integrated promoter prediction (*GenieESTPROM*) indicates a possible TSS at 1,111,271 for the *Adhr* gene with a reasonable score. It would be interesting to verify this prediction by biological experiments.

Figure 5-12: *Adh-Adhr.* Annotations for the following known genes described in Ashburner *et al.* are shown for the region from 1,109,500 - 1,112,500 (forward strand only) (from the left to the right of the map):
***Adh, Adhr*.**



Analysis of the gene *outspread* (*osp*), Figure 5-13, reveals a structural shortcoming in the gene model of Genie. The *outspread* gene, the first gene on the reverse strand, contains many very long introns and contains within one of these introns the *Adh*/*Adhr* gene pair on the opposite strand. In another intron *outspread* contains one gene (*DS09219.1*) on the same strand and another one like *Adh/Adhr* on the opposite strand (*DS07721.1*). The current Genie model is built in such a way that it does not allow gene(s) within or overlapping other genes either on the same or on the reverse strand. Therefore Genie breaks up the *outspread* gene. The seven 3' exons are predicted correctly, but Genie introduces an erroneous first exon to complete this gene prediction. The exon at 1,104,419 - 1,104,995 overlaps with a Genie prediction of a single exon gene from 1,104,411 to 1,104,965. The correct prediction of most protein coding bases in *outspread,* despite the program's inability to identify the full gene structure in this complex situation, demonstrates its graceful degradation on unusual gene structures and may explain its high base-level sensitivity relative to the number of totally correct gene predictions. While the remaining seven 5' exons from *outspread* are missed, the *GenieEST* and *GenieESTHOM* versions introduce a wrong three-exon gene in the middle of an intron. These EST-based Genie versions are forced to predict this gene through a mistaken EST sequence hit and alignment, which belongs to the overlapping *DS09219.1* gene transcript (see Table 9 for details).

Figure 5-13: *outspread.* Annotations for the following known genes described in Ashburner *et al.* are shown for the region from 1,090,000 - 1,180,000 (from the left to the right of the map):
***outspread* or *osp* (r), *Adh* (f), *Adhr* (f), *DS09219.1* (r), *DS07721.1* (f).**



Additional evidence for the general Genie behavior of splitting genes comes from the most complex gene in the *Adh* region, the *Ca-alpha 1D* gene (Figure 5-14). This long gene with more than 30 exons is incorrectly split by Genie into three separate genes. Most of the long exons are covered by Genie predictions, but some of the short exons are missed entirely.

Figure 5-14: *Ca-alpha1D.* Annotations for the following known gene described in Ashburner *et al.* are shown for the region from 2,617,500 - 2,640,000 (forward strand only) (from the left to the right of the map):
***Ca-alpha1D*.**

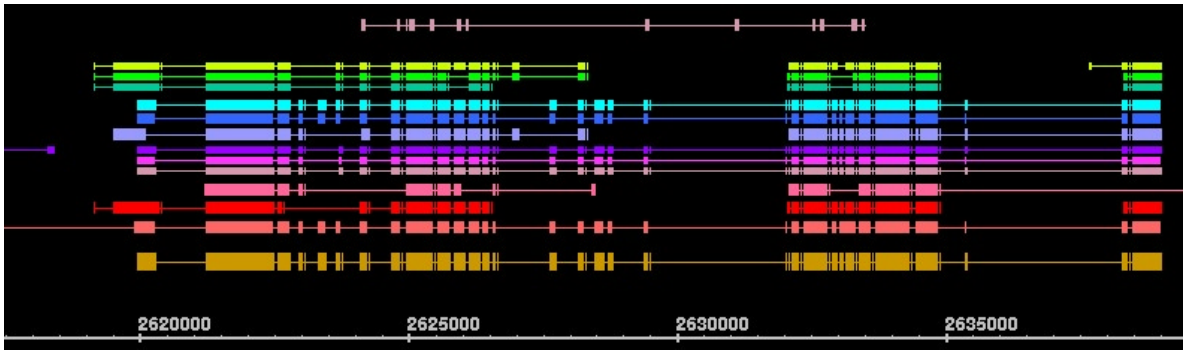
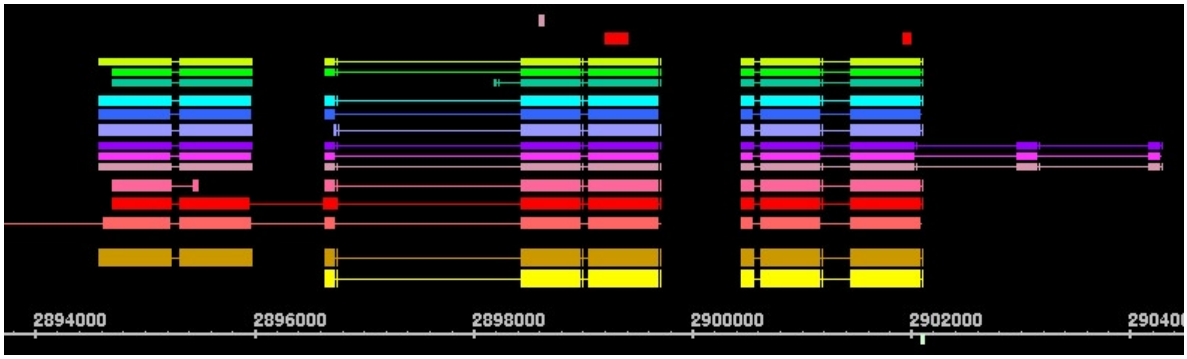


Figure 5-15, the *idgf* cluster of three genes of the same family (*idgf1, idgf2,* and *idgf3*), shows the benefit of using EST information and the additional benefit of using homology information in Genie. The first intron is missed by the statistical *Genie* but recovered through the additional EST alignment information that spans this intron in

GenieEST. *Idgf3* is correctly predicted, but *GenieESTHOM* only entirely correctly predicts *idgf1*. For *idgf1* the protein homology information extends the initial exon to a different start codon further upstream.

Figure 5-15: *idgf* cluster. Annotations for the following known genes described in Ashburner *et al.* are shown for the region from 2,894,000 - 2,904,000 (forward strand only) (from the left to the right of the map):
***idgf1, idgf2, idgf3*.**



5.4 Additional selected observations of the Genie annotation

In Table 5-2 twenty-six potential novel genes are listed that are predicted by at least one of the Genie programs and in addition have evidence through an overlap from at least one other gene finding or homology program. The number seems to be very high (over 10.1% of a total number of 258 genes predicted from GenieESTHOM), but because the process of annotating genes in genomic DNA is so hard, and not all programs were available at the time of the annotation (Ashburner et al., 1999), I believe that at least the majority of these predictions are real genes. All predictions that overlap with an annotated transposable element were removed from this list.

Table 5-2: Predicted novel genes by Genie. Twenty-six Genie gene predictions that have no overlaps to any gene structure in *std3* are listed. "Strand" indicates the strand on which the predicted genes are located. The strand is consistent with the annotation in Ashburner *et al.*: "Genes on the top [forward strand] of each map are transcribed from distal to proximal (with respect to the telomere of chromosome arm *2L*); those on the bottom [reverse strand] are transcribed from proximal to distal". "Begin" and "End" gene coordinates note the first and last base of the predicted coding gene region by Genie. *Genie*, *GenieEST* and *GenieESTHOM* label the Genie program variant. "Other gene finder hits" lists the count of how often this newly predicted gene is also overlapped by one or more of the other six gene finding programs from GASP. "Homology hits" marks the count of how often a newly predicted gene overlaps any homology hits from GASP.

Strand	Begin	End	Genie	Genie EST	Genie EST HOM	Other gene finder hits	Homology hits	Comments
F	21,599	21,988	X	-	-	5	0	
F	131,015	131,248	X	X	X	5	0	
F	267,633	268,061	X	X	X	1	0	
R	306,476	306,985	X	X	X	1	0	
R	328,048	328,733	X	X	X	4	0	
F	329,808	331,184	X	X	X	6	0	
F	403,468	405,391	X	X	X	5	2	
F	408,759	412,000	X	X	X	6	2	
R	426,746	427,525	X	X	X	6	0	
R	603,442	604,456	-	X	X	5	0	
F	754,773	754,919	X	X	X	2	0	
R	846,339	845,892	-	X	X	3	0	
R	870,684	870,866	X	X	X	4	0	
R	910,572	911,055	X	X	X	5	0	
F	1,115,807	1,116,493	X	X	X	2	0	
F	1,117,474	1,117,608	X	X	X	3	0	
R	1,263,535	1,264,137	-	X	X	3	0	
R	1,365,077	1,365,732	X	X	X	4	0	
R	1,850,650	1,851,240	X	X	X	4	0	
F	2,453,955	2,454,498	-	X	X	4	0	
F	2,580,916	2,581,059	X	X	X	1	0	FP (?)
R	2,584,165	2,584,914	-	X	X	3	0	
R	2,741,387	2,742,230	X	X	X	4	0	
R	2,762,639	2,774,287	X	X	X	2	0	
F	2,779,268	2,779,566	X	X	X	2	0	
F	2,843,324	2,843,386	X	X	X	1	0	FP (?)

Table 5-3 lists the nineteen genes from the reference *std3* set for which no overlap of a Genie prediction exists. Thus, less than 10% of the annotated genes in *std3* are missed by Genie. The individual submission scores for missed genes are as low as 4.6%. Nine of these *std3* annotations are solely based on predictions from the human version of GENSCAN (Burge & Karlin, 1997) and/or Genefinder (Green, 1995) predictions, the two gene finding programs used for the annotations in Ashburner *et al.*. For an additional six of these genes Ashburner *et al.* augmented their evidence with BLAST hits with low P-values. For *Mst35Bb* there exists a very reliable cDNA alignment, it is definitely a real missed gene. The remaining three genes (*DS07721.1*, *DS003192.3*, *DS003192.4*) are all based on cDNA alignments. None of them is predicted by any gene finding program and for one, *DS003192.4*, there are two homology annotations but on the opposite sequence strand! Therefore these alignments are very questionable and might be the result of typical cDNA cloning artifacts.

To summarize the analysis of the nineteen over-predicted genes it is possible that the missed gene prediction rate of Genie is below the noted 4.6% and that very few real genes are missed. This is believed to be also true for some of the predictions of the other programs.

Table 5-3: Genes missed by Genie. The gene names from Ashburner *et al.* (1999) are listed. In addition, the evidence for that gene annotation in that paper is given. The "Begin" and "End" gene coordinates are from the *std3* annotations. In addition, the number of overlaps by other gene finders and the two homology programs are listed.

Strand	Begin	End	Other gene finder hits	Homology hits	Gene names in Ashburner et al. (1999)	Evidence in Ashburner et al. (1999)	Comments
R	230,985	240,152	3	0	DS08249.5	Gene pred. only	-
F	498,520	507,581	2	0	DS01514.3	Gene pred. only	-
R	523,395	525,283	3	0	DS05899.7	Gene pred. and low P- value BLAST hit	-
R	533,592	536,913	3	0	DS05899.6	Gene pred. only	-
F	1,152,128	1,152,385	0	0	DS07721.1	cDNA (?)	Suspicious annotation
R	1,285,030	1,286,199	3	1	DS06874.6	Gene pred. and low P-value BLAST hit	-
F	1,300,469	1,315,922	3	0	DS06874.7	Gene pred. only	-
R	1,368,793	1,369,282	5	0	Mst35Bb	cDNA	-
F	1,484,701	1,489,834	2	0	DS00929.16	Gene pred. only	Suspicious annotation
R	1,520,808	1,521,371	1	1	DS00929.7	Gene pred. and low P-value BLAST hit	Overlaps gene on opposite strand
F	1,628,242	1,628,412	0	0	DS003192.3	cDNA (?)	Very suspicious annotation
R	1,663,026	1,663,163	0	(2 oppo. strand)	DS003192.4	cDNA (?)	Very suspicious annotation
R	1,782,412	1,786,409	2	0	Ms(2)35Ci	Gene pred. only	-
R	1,875,987	1,895,879	5	0	DS03023.4	Gene pred. only	Gene on opposite strand
R	2,109,315	2,113,209	4	0	BACR44L22	Gene pred. only	-
F	2,158,476	2,159,460	3	2	DS07108.5	Gene pred. and low P-value BLAST hit	-
F	2,236,081	2,241,876	3	0	DS02252.3	Gene pred. and low P-value BLAST hit	5,000 bp single exon gene
R	2,286,435	2,287,433	3	0	DS02252.4	Gene pred. and low P-value BLAST hit	-
F	2,398,367	2,410,394	5	0	DS07486.5	Gene pred. only	-

One of the biggest problems with the Genie programs in annotating *Adh* was joined and split genes. Table 5-4 and Table 5-5 show that Genie is parameterized to favor splitting genes versus joining genes. Only nine Genie annotations span two or more *std3* genes (*Joined genes*) while nineteen *std3* genes are split into separate Genie predicted genes (*Split genes*). The problem of joining genes is due to the difficulty of identifying the ends and starts of genes that don't encompass strong statistical signals. Careful analysis of the split genes, on the other hand, showed that the length distribution of introns - a geometric distribution - which favors short introns, is the reason for so many split genes (data not shown). Another behavior related to the same problem of the length distributions of introns is the general tendency of Genie to introduce erroneous exons within otherwise long introns. Table 5-6 lists eleven typical examples.

Table 5-4: Joined genes by Genie. All predictions in which Genie joins one or more genes from *std3* are listed. The "Begin" and "End" gene coordinates are from the Genie predictions. The last two columns list the number of genes joined and their respective names.

Strand	Begin	End	Genie	Genie EST	Genie EST HOM	# of joined genes	Names of joined genes
R	336,668	339,013	X	X	X	2	*DS00941.11, DS00941.12*
F	341,713	343,984	X	X	X	2	*DS00941.14, DS00941.15*
F	454,701	458,802	X	X	X	2	*DS00180.5, DS00180.12*
R	458,837	463,657	X	X	X	2	*DS00180.7, DS00180.8*
F	471,109	476,389	X	X	-	2	*DS00180.11, DS00180.14*
R	839,712	843,808	X	X	X	3	*DS01068.10, DS01068.4, DS01068.5*
F	1,599,218	1,607,306	X	X	X	2	*DS04929.3, stc*
R	2,102,169	2,104,442	-	-	X	2	*BACR44L22.8, BACcr44L22.2*
R	2,786,019	2,792,601	X	X	X	3	*DS02740.18, DS02740.19, DS09218.1*

Table 5-5: Split genes by Genie. All *std3* gene annotations that are split into two or more genes by all three Genie programs. The "Begin" and "End" gene coordinates are from the *std3* annotations. The causes of splitting, when known, are noted in the Comments column.

Strand	Begin	End	# of split genes	Comments
F	45,358	130,409	5	Gene on opposite strand
F	373,286	391,500	3	
F	445,189	456,317	3	
R	477,171	487,236	2	
F	568,986	575,533	2	
R	679,874	691,416	2	
F	757,457	821,487	4	1 gene on same strand
R	1,094,414	1,182,415	2	2 genes on opposite strand
R	1,398,183	1,413,067	2	
F	1,506,022	1,521,842	2	Gene on opposite strand
F	1,558,915	1,561,694	2	
F	1,565,296	1,585,380	3	
R	1,653,146	1,667,970	2	
R	1,718,580	1,737,780	2	
R	1,747,063	1,752,780	2	
R	2,220,563	2,224,367	2	
F	2,463,394	2,488,789	2	
F	2,619,967	2,639,006	3	
R	2,714,362	2,736,449	2	

Table 5-6: Missed long intron(s) by Genie. Genes that have long introns that are missed by any Genie program are listed and it is indicated which program misses them. The "Begin" and "End" coordinates are from the *std3* annotations.

Strand	Begin	End	Genie	Genie EST	Genie EST HOM
R	268,751	273,483	X	X	X
R	654,984	667,105	X	X	X
F	828,047	833,672	X	X	X
R	880,856	901,495	X	X	X
R	1,051,748	1,057,314	-	-	X
R	1,271,377	1,276,359	X	X	X
R	1,421,921	1,432,223	X	X	X
F	1,974,488	1,983,855	X	X	X
F	2,040,123	2,057,901	X	X	X
F	2,505,534	2,530,156	X	X	X
F	2,683,427	2,694,719	X	X	X

A simple but serious oversight in the GASP experiment the poor treatment of transposable elements in the *Adh* region. Ashburner *et al.* found seventeen transposable elements, which consist of repetitive elements but also include protein coding like regions, including long open reading frames, predominantly for the transposase and the reverse transcriptase proteins. As expected, Genie cannot distinguish these transposon genes from protein coding genes and therefore predicts thirteen of the existing seventeen as protein-coding genes (see Table 5-7 for a list of predicted transposons). In particular, *GenieESTHOM* predicts many of the transposable elements to be coding genes because transposable elements contain protein sequences that result in strong protein alignments. While the statistical *Genie* version only overlaps three of the seventeen transposable elements, *GenieESTHOM* predictions overlap thirteen. These transposon hits contribute to an increased false positive rate, worse *wrong exon* and *wrong gene* scores, and lower overall specificity in Table 5-1.

Table 5-7: Transposable elements. Transposable elements incorrectly labeled as real gene by Genie. The transposable elements that have an overlapped prediction by Genie are listed. The "Begin" and "End" coordinates are the transposable element coordinates from the Ashburner *et al.* (1999) paper.

Strand	Begin	End	Genie	Genie EST	Genie EST HOM	Other gene finder hits	Homology hits	Transposon name
F	55,422	58,941	-	-	X	4	2	*Fw*
R	93,549	94,119	X	X	X	3	1	*G*
R	255,612	256,662	-	-	X	1	1	*Doc*
R	959,378	962,797	-	-	X	2	1	*Doc*
R	1,136,806	1,145,466	-	X	X	5	0	*Roo*
R	1,293,597	1,298,741	-	X	X	5	1	*Copia*
F	1,474,114	1,481,634	X (2 genes)	-	-	3	2	*Yoyo*
F	1,935,760	1,943,170	X	X	X	3	2	*Blood*
F	2,076,116	2,083,110	-	-	X	3	2	*297*
F	2,174,330	2,176,188	-	-	X	1	1	*Copia-like*
F	2,177,045	2,178,655	-	-	X	3	2	*Copia-like*
F	2,590,477	2,595,625	-	-	X	5	2	*Copia*
F	2,603,050	2,610,046	-	-	X	2	1	*297*

In Table 5-8 five gene annotations are reported based on Genie predictions that strongly indicate either a different gene structure than reported in *std3* or a potentially new alternative splicing form for the listed genes. The underlying evidence, besides the Genie predictions, comes from other gene finding predictions as well as from EST sequence alignments.

Table 5-8: Alternative splicing forms predicted by Genie. Genes in *std3* that might have an alternative gene structure as predicted by Genie and other gene finders are listed.

Strand	Annotations	Begin	End	Genie	Genie EST	Genie EST HOM	Other gene finder hits	Gene name	Comments
F	*std3*	159,578	163,527	X			4	*DS01368.1*	EST alignment verifies last additional intron
	Genie	159,578	164,417	X	X	X	5		
R	*std3*	325,240	326,379					*MtPolB*	Additional first exon
	Genie	325,240	326,822	X	X	X	3		
R	*std3*	1,334,780	1,338,785	X			5	*DS03431.1*	Missed third exon, (EST verified)
	Genie	1,334,780	1,338,785		X	X	1		
R	Std3	1,371,813	1,372,351				2	*Mst35Bb*	Longer 1st exon and shorter last exon
	Genie	1,371,868	1,372,213		X	X	3		
F	*std3*	1,493,680	1,496,198				1	*DS00929.1*	Wrong first exon and EST intron in 2nd exon
	Genie	1,495,484	1,496,198	X	X	X	4		

Through evidence from high scoring Genie predictions, EST alignments and the other GASP annotation teams, eight gene entries in the *std3* reference set seem to be very suspicious. In Table 5-9, predictions from other programs are only listed if they support the suggested corrected gene structure annotated by Genie. Careful cDNA alignment and additional full-length cDNA sequencing should shed light on these cases in the future.

Table 5-9: Possible "incorrect" annotations from *std3*. Genes from the *std3* annotations are listed, for which multiple evidence from Genie and other programs exists, implying "incorrect" annotations. The "Begin" and "End" coordinates are from the *std3* annotations. The evidence for the annotation in *std3* is given as noted in Ashburner *et al.*.

Strand	Begin	End	Genie	Genie EST	Genie EST HOM	Other	Homology hits	Gene name	Evidence in Ashburner *et al.*	Evidence from predictions
R	213,507	217,188	X	X	X	7	1	*DS08249.3*	Gene pred. only	Last exon questionable
F	281,649	284,052	X	X	X	6	2	*D00797.5*	cDNA (partial?)	10 leading exons missing
R	941,115	944,598	X	X	X	4	-	*DS08340.1*	Gene pred. only	4 extra 3' UTR exons
F	1,205,439	1,213,325	X	X	X	5	-	*DS07721.3*	Gene pred. only	First exon and last 3 exons questionable
R	1,371,813	1,372,351	-	X	X	3	-	*TFIIS*	Known gene	Longer first and shorter last exon
R	1,549,142	1,549,933	X	X	X	6	-	*DS07295.4*	Gene pred. only	Initial exon and first intron verifired by EST missing
F	1,721,863	1,728,736	X	-	-	1	-	*DS07295.4*	Gene pred. only	At least first 7 exons very questionable
R	1,913,374	1,914,948	X	X	X	6	1	*wor*	Gene pred. and BLAST homology hits	Additional first exon verfied by EST

5.5 Promoter prediction results in Genie

A total of 234 transcription start site predictions were produced by the integration of NNPP into the statistical Genie version, *GeniePROM,* and 237 TSS's were predicted by the EST refined version of Genie, *GenieESTPROM*. The success rate of the promoter assignment of about 30% (27.6% for *GeniePROM* and 32.9% for *GenieESTPROM*) is in the same order as other promoter predictions from GASP, but indicates that promoter recognition is very difficult due to the complex initiation process (see **Table 5-10** taken

from Reese *et al.* (2000)). Because Genie's promoter assignments are in the context of gene identification and as such are modeled in the complete generalized HMM to occur upstream of the start codon, the false positive rate is low. For the evaluation of 856,119 negative bases the rate is 1/14,760 for *GeniePROM* and 1/16,786 for *GenieESTPROM*, respectively. It is interesting to recognize that the EST integration improves promoter identification, which might be due to an extension of the 5' region of a gene using information from a 5' EST sequence. Because of the integration into a gene finding system the numbers should be compared with the similar MAGPIE system and it can be seen that while *GenieESTPROM* misses two more promoters (31 versus 33) its false positive rate is lower (1/16,786 versus 1/14,760).

In addition, in Table 5-10 the prediction statistics for the pure NNPP program are added without the integration in Genie with thresholds of 0.97 and 0.90. At a threshold of 0.97 the number of 35 identified TSS's is very similar to the Genie integrated system GeniePROM of 30 TSS predictions, but the false positive rate is seven times higher - 1/2,416 bp versus 1/14,710 bp. If one is interested in getting as many correct predictions as possible the only solution is to run NNPP with a low threshold (t=0.9) to predict more than 50% of the real TSS's. Neither GeniePROM nor GenieESTPROM nr any other program tested in GASP is able to predict close to this number. Nevertheless, the high specificity of the promoter prediction systems integrated in gene finding systems is obviously due to the context information: all promoter predictions within gene predictions are ruled out in advance, and the location of the possible start codon provides the system with a good initial guess as to where to look for a promoter.

Table 5-10: GASP evaluation of promoter prediction programs. We show the number and percentage of identified transcription start sites in comparison to the false positive rate which is given for two different sets of regions: (a) the "likely" region for a transcription start site plus the downstream region belonging to the same annotation; (b) the same region plus half the distance to the neighboring genes upstream and downstream (taken from the *std3* annotation).

System Name	Identified TSS	Rate of false predictions in region (a) (853,180 bases)	Rate of predictions in region (b) (2,570,232 bases)
NNPP (t=0.97)	35 (38.0 %)	1/2,416	1/1,019
NNPP (t=0.90)	55 (53.2 %)	1/928	1/404
CoreInspector	1 (1 %)	1/853,180	1/514,046
MCPromoter V1.1	26 (28.2 %)	1/2,633	1/2,537
MCPromoter V2.0	31 (33.6 %)	1/2,437	1/2,323
GeniePROM	25 (27.1 %)	1/14,710	1/28,879
GenieESTPROM	30 (32.6 %)	1/16,729	1/29,542
MAGPIE	33 (35.8 %)	1/14,968	1/16,370

5.6 Genie improvements after GASP

The excellent test of Genie in the GASP experiment helped further improvement and fine-tuning of the system to produce even more reliable gene annotations for the entire genome of *Drosophila melanogaster*.

The oversight of the non-treatment of transposable elements that resulted in many false positive predictions in each performance category – many coding regions in transposable elements were mistaken as genes, especially when using protein homology – was corrected by a simple pre-screening method for transposable elements. This prescreening masked out these transposon regions and eliminated them from being predicted by Genie.

Another structural mistake in the EST based Genie gene models, *GenieEST* and *GenieESTHOM*, resulted in erroneous predictions when EST evidence identified introns between non-coding exons (Table 5-11 gives a list of these false predicted genes in

Adh). This happened because Genie's exon and intron models were exclusively based on coding region of genes and any predicted 5' or 3' UTR intron from an EST alignment was mistakenly predicted as a coding exon. For the Genie application to the entire genome the underlying GHMM gene model was changed by adding the notion of an intron in an UTR region.

Table 5-11: Erroneous EST UTR predictions by Genie. Coding gene predictions by *GenieEST* and *GenieESTHOM* that are either complete over-predictions or partially wrong by extending the coding regions into the 5'/3' UTR due to a wrong underlying gene model structure for non-coding ESTs (see text for details). The "Begin" and "End" coordinates are the *GenieEST* and *GenieESTHOM* predictions.

Strand	Begin	End
R	40,843	43,076
R	346,994	356,311
R	393,573	398,794
F	507,364	512,758
F	849,268	851,919
R	1,372,338	1,373,546
R	1,756,026	1,761,674
F	2,491,469	2,497,464
R	2,698,932	2,706,347
R	2,709,485	2,711,209

Another improvement in the underlying gene model allows it to combine gene information from 5' and 3' EST sequences that were sequenced from the same cDNA clone - hopefully a full-length clone. This knowledge can be used in gene finding to restrict the gene boundaries: the gene start and end. This was modeled by constraining the GHMM framework to refuse to predict intergenic regions between a 5' and 3' aligned EST sequence from the same cDNA clone.

Chapter 6 Discussion

In this thesis I have described two computational methods both rooted in the field of machine learning. I have used these methods to model structural and sequence composition properties of the *Drosophila melanogaster* genome and I have applied both of these models to the problem of transcription start prediction and gene identification in the complete genome of *Drosophila melanogaster*.

The first tool is an artificial neural network model using a time-delay network architecture. This network has two feature layers: one for the TATA box and one for the *Inr* (initiator). The output of both feature layers is combined in a time-delay neural network. I have shown that such a neural network detects the TATA box and the Inr and is insensitive to their relative spacing and is therefore an excellent model for the compositional sequence properties of a eukaryotic core promoter region. The discriminative ability of such a model for the short core promoter region of -40 to +11 bases spanning the transcription start site is so strong that this model can be used to predict an entire promoter in genomic DNA. These results show that the highest information content in a promoter region exists in the core promoter region.

The NNPP computer program implements the time-delay neural network model. The program is able to predict over 70% of transcription start sites in genomic DNA when used with the default parameters. The false positive rate calculated on the *Adh* region *in Drosophila melanogaster* is 1/ 547 bases. The Matthew's correlation coefficient (Matthews, 1975) is 0.58. 30% of all promoter sequences remain undetected and this is probably due to the non-local structure of the promoter region, where initiation control elements can occur at positions many kilobases distant from the transcription start site. In a published comparison of eukaryotic promoter prediction tools (Fickett & Hatzigeorgiou, 1997) the NNPP program performed better than other similar programs.

The NNPP program can easily be extended to incorporate novel information as it becomes available. Other known promoter elements such as the CAAT box, GC box, DPE (downstream promoter element; so far known to exist only in *Drosophila*), and conserved transcription factor binding sites can also be used within the existing framework. The extended parameter space of such an extended model would require more data for training.

The positive results obtained using the time delay architecture will hopefully lead to more widespread application of neural networks to similarly complex problems in molecular biology, such as the detection of splice sites and protein-protein interaction motifs.

For the application to complete genome annotations the NNPP code is integrated into the Genie system as described in Section **4.4.3**. Such integration is necessary to reduce the false positive rate that naturally occurs by modeling complex systems like promoters. Results of the integrated program with standard setting show a recognition rate of 32.6% with a more "realistic" false positive rate of 1 false prediction in 16,729 bases.

Since I made the NNPP program available on the World Wide Web it has been widely used in the scientific community to hypothesize about potential transcription start sites. One of many applications of NNPP is presented as an example (experimental results provided by Roehrig (2000), personal communication):

The *C. elegans* gene *unc-86* encodes a POU IV class transcription factor, which is expressed exclusively in the nervous system. There are currently no ESTs available for *unc-86*, since the gene appears to be expressed at low abundance (the mRNA is not detectable in Northern blot analysis). This has hampered the identification of potential alternative splice sites and the transcription start sites. The *ab initio* gene finder "Genefinder" (Green, 1995) does not predict the 5' region of the CDS correctly. 5'-RACE studies (Roehrig, 2000) revealed that the first exon codes for only three amino acids, making its *ab initio* prediction very difficult. Yet, it remained uncertain whether the amplified products in the 5'-RACE truly represented the 5'-end of the mRNA. In order to address this problem, the *unc-86* genomic sequence was analyzed *in silico* using NNPP.

NNPP predicted the transcription start site to be approximately 120 bp upstream of the site identified in the 5'-RACE experiments. This predicted TSS would allow translation to start at an ATG codon 45 bases upstream of the TSS previously predicted by RACE. This would add another 15 amino acids to the N-terminus of the *unc-86* protein. This prediction was experimentally supported by the amplification of a cDNA that included the additional 5' region predicted by NNPP. In parallel, sequence analysis from the closely related nematode species *C. briggsae* revealed that the DNA sequence encoding the putative additional 15 amino acids is conserved, while further upstream of the NNPP-predicted TSS, the sequence is not conserved. Finally, primers were selected from the sequence upstream of the NNPP predicted TSS and they failed to amplify a cDNA product in RT-PCR experiments (Roehrig (2000), personal communication).

Both Genefinder and Genie predicted the remaining 3' gene structure in accordance with the experimental cDNA alignment data. Sequence comparison with the *C. briggsae* sequence does not hint at alternatively spliced exons (Roehrig (2000), personal communication).

This example demonstrates how useful a program like NNPP can be the right context. It is clear that a program cannot substitute for the final experimental proof but the example shows that it can give direction and guidance for such experiments to verify computational predictions.

I believe *Drosophila melanogaster* is a good model organism to study transcription. To get a better understanding of transcription in general, better and more complete data sets are needed. The most comprehensive collection of promoters, the EPD database (version 61_1), contains a collection of only 807 experimentally verified non-redundant eukaryotic POL II promoters, 108 of which were found in *Drosophila*. Initial efforts at the BDGP and at Harvard University have extended this set to a total of 265 (Ohler, 1999) annotated promoters with experimentally verified TSS's. This is still a very small number considering a predicted total number of at least 13,000 *Drosophila* genes (Rubin *et al*., 2000). It means that for only 2.04% of the total genes there exist experimentally verified and annotated promoters. An extended set could be generated using for example, the RACE protocol above described in combination with *in silico* methods such as NNPP.

The second tool presented in this thesis is a probabilistic model of the gene structure in *Drosophila*. The main accomplishment was the development of a new gene identification program specifically for *Drosophila*, called Genie, which has several significant advantages over existing gene finding programs. Most importantly, the accuracy of Genie is higher than for any other available method when tested on large genomic sequences such as the *Adh* sequence, as was demonstrated in the GASP experiment. Genie shows a sensitivity and specificity at the nucleotide level of 97 % and 92%, respectively. It predicts splice sites and start and stop codons with high confidence. For example, the program identifies 79% of exons in genomic sequence accurately. This high accuracy is the result of the ability to use cDNA sequence alignments to verify the exact boundaries between exons and introns and between gene assemblies.

The Genie system successfully integrates a combination of *ab initio* gene finding signal and content models, EST/cDNA alignments, protein homology alignment information and promoter signal models specific to the *Drosophila* genome. Genie is most similar to GENSCAN (Burge & Karlin, 1998) in its overall architecture. It is, like GENSCAN, a probabilistic model that uses a generalized hidden Markov model of gene structure (in the GENSCAN literature this is called “semi-hidden” Markov model). Initial developments of both programs were done very much in parallel and were driven by sharing concepts and training as well as testing data. While GENSCAN includes submodels for poly-adenylation sites and signal peptides, these signals are not integrated in Genie, because they did not show a statistical significant improvement in accuracy. The main difference between GENSCAN and Genie is the integration of EST sequence alignments and protein homology information in Genie. Furthermore, we performed extensive parameter optimization specifically for use in *Drosophila*. The optimization included, among other things, collecting additional *Drosophila* mRNA sequences, adjusting for the specific intron length distribution and pre-masking for known transposable elements. Due to the lack of the availability of a *Drosophila* version of the GENSCAN program, performance accuracy could only be compared to an older version as part of the MAGPIE submission for the GASP experiment. This comparison showed a significant improvement of the GENSCAN results by Genie.

The limited accuracy of gene assembly is problematic in Genie's model, and I believe it is true for most of the gene finding systems. While the base level predictions

and exon level predictions are very good, the results for gene assemblies are not (this is partially reflected in the scores for *joined* and *split genes*). Because the GASP experiment showed that splitting genes was a major problem, the generalized Hidden Markov Model (GHMM) framework was extended to integrate information about the pairing of 5' and 3' EST sequences from the same clone. This change improved gene assembly but is of no use if no EST coverage exists (at the current time the total gene coverage in the *Drosophila* EST database of paired 5' and 3' ESTs is estimated to be between 50-60% (See the BDGP EST project)).

A second limitation of the Genie system is the inability to detect genes within the introns of other genes. The limited number of examples of *Drosophila* overlapping genes has been a significant impediment. It will be interesting to see how EST alignment information might be helpful for predicting genes within genes but it still remains a challenging problem.

In order to fully assess the Genie predictions and in particular the 26 novel predictions in the *Adh* region discussed in Section **5.4**, the "correct answer", in the case of the GASP experiment the annotations in the *Adh* region, must be improved. Only extensive full-length cDNA sequencing can accomplish this. A possible approach would be to design primers from predicted exons/genes in the genomic sequence and then use hybridization technologies to screen for the corresponding cDNA from multiple cDNA libraries. Initial experiments are underway to verify these additional *ab initio* predictions from the *Adh* region as a result of GASP experiment (Rubin, 2000).

Genie was developed to be used for gene identification in the annotation process of the entire genome of *Drosophila melanogaster.* In collaboration between the Berkeley Drosophila Genome Project and Celera, Inc., the genome of *Drosophila melanogaster* has been sequenced (Adams *et al*., 2000; Rubin et al., 2000). GenieEST, the best performing version of Genie in GASP, was successfully run on a 10-fold sequence assembly of the complete genomic sequence. In total, Genie predicted 13,189 genes, while the other *ab initio* program GENSCAN (see discussion above (Burge & Karlin, 1997)) predicted 17,464 genes (Rubin et al., 2000). Based on the results from the GASP experiment, where GENSCAN was used as a part of the MAGPIE system, we believe that the lower number predicted by Genie is more accurate. In *Adh,* Ashburner *et al.* (1999) annotated a total of 218 genes, which was later adjusted to 222 genes in a second

round of analysis during GASP. GENSCAN predicted 468 genes while Genie predicted only 246 genes for this region. The main differences are discussed above. The GASP experiment showed that due to the non-optimized parameters in GENSCAN, more over-predictions occur, and that the Genie predictions are closer to the experimentally verified number of 222.

Extrapolating from the *Adh* region to a complete genome size of 110 Mbases gives a very conservative estimate of 9,293 total genes. The difference between this lower number and the total predicted number of 13,189 might be due to sequencing errors, sequence gaps in the 10 fold-assembly or an atypical lower gene frequency in the *Adh* region.

Preliminary results from observations during the annotation process for the *Drosophila* genome in November 1999 show that the Genie predictions are very reliable and can very often be verified by additional sequence alignment information from ESTs and homologous proteins. Additional evidence for the accuracy and completeness of Genie is given by the observed high percentage of expected genes for Drosophila that were found in the predicted protein set by Genie (see Rubin *et al* (2000) for more details).

Chapter 7 Conclusion

The goal to develop an accurate and robust system for the identification of *Drosophila* genes and to successfully apply the system in the annotation process of the complete genome of *Drosophila melanogaster* has been achieved. Furthermore, we built a powerful program for predicting and studying transcription start sites.

Over the years, Genie has become a robust gene finding system. It is highly modular and allows for automatic training for new organisms, and the integration of new external sensor models; It runs fully automatically for entire genomes and the running time is reasonable when applied to complete genomes such as the human genome. The statistical framework allows for a probabilistic assessment of individual predicted features and complete gene predictions. The concept of a generalized hidden Markov model first introduced in early Genie related publications (Kulp et al., 1996; Reese et al., 1997), is very powerful, as can be seen in the high performance scores of all systems based on GHMMs in the GASP experiment.

The GASP experiment provided an objective assessment of current approaches to gene prediction, which proved to be very useful for the final complete genome annotation process. The main conclusions from the experiment are that current methods of gene predictions have tremendously improved and that they are very useful for genome scale annotations. However, high quality annotations also depend on a solid understanding of the organism in question (*e.g.,* recognizing and handling transposons). Experiments such as GASP are essential for the continued progress of automated annotation methods. They provide benchmarks with which new technologies can be evaluated and selected.

Beyond the identification of gene structure is the determination of gene function. Most of the existing prototypes for such systems are based on sequence homologies.

While this is a good starting point it is definitely not sufficient. The state of the art for predicting function for protein sequences uses the protein's three-dimensional structure, but the difficulty of accurately predicting three-dimensional structure from primary sequences makes applying these techniques on complete genomes problematic. The new field of structural genomics will hopefully give more answers in these areas.

In summary, both models and their implementations, NNPP and Genie, can be considered a step forward towards the goal of identifying all *Drosophila* genes. In a more general context they can be viewed as tools that empower scientists to reach a better understanding of the fundamental complex biological processes involved in gene regulation and gene localization.

Chapter 8 Appendices

Appendix A URLs

Gene data sets: http://www.fruitfly.org/sequence/drosophila-datasets.html

http://www.fruitfly.org/sequence/human-datasets.html

NNPP: http://www.fruitfly.org/seq_tools/promoter.html

Genie: http://www.fruitfly.org/seq_tools/genie.html

NNSplice: http://www.fruitfly.org/seq_tools/splice.html

GASP: http://www.fruitfly.org/GASP1

Appendix B Promoter data sets

429 unrelated eukaryotic promoters taken from the Eukaryotic Promoter Database (EPD;(Bucher, 1990) release 41. The EPD accession numbers are listed.

BTBPTIG1_16066	BTHOR01_07102	BTKER4_15028	BTKER6B_11082	BTKERAIB_15027	BTKERIA1_15026
BTPGPHA1_11126	BTPPT1_11122	BTPROB_28006	BTPTHG_30046	BTSIG1_16067	CCTPMY01_25016
CMHIST34_33027	CMHIST34_33028	FSPRC2A_17043	GDCOLG2_30050	GDCTNT_24007	GDHMG141_31008
GG5ACT1_11079	GGACHRA_15042	GGACTAC_11078	GGACTI_07059	GGAL07_07081	GGALASY1_14048
GGALDB_17060	GGC1A201_07066	GGCAII12_23006	GGCALB_07085	GGCAMP_24005	GGCRYDS_11086
GGFERH_16048	GGGADPHE_30015	GGGHRA05_30049	GGGL02_07075	GGGL03_33032	GGGL04_11095
GGH2A2B_33016	GGH2A2B_33017	GGH2AF_33018	GGH2B1_24013	GGH33B_33020	GGHBBR2_07076
GGHI03_11067	GGHIS1_07050	GGHISH1_11065	GGINS1_07110	GGKERC_07068	GGLYSX_07087
GGMY04_07064	GGMYC_15048	GGMYHE_11081	GGOV03_07082	GGOVO1_07086	GGP4HB01_30040
GGPGR_25037	GGRIBPRL5_41004	GGRIG_37012	GGRPL30_41005	GGRPL37A_41006	GGTIMA_30016
GGTROSS1_25018	GGU4BX_17040	GGVI01_07089	GGVIM1_25002	GGVL01_07090	HGGL01_07070
HSA1ATCA_17090	HSA1ATP_17092	HSABL1B_26030	HSABLA_26032	HSACTBPR_17045	HSADAG1_11113
HSADPRF01_41009	HSADPRF1_39001	HSALBEX1_16042	HSALDA1_26015	HSALDA1_26017	HSALDB1_11119
HSALDH01_23004	HSAMYAGA_30065	HSANFG1_11132	HSAPB01_26028	HSAPOA2_11088	HSAPOAI1_30021
HSAPOC2G_17051	HSAPOE4_36007	HSARG1_30054	HSASG5E_11114	HSATH2_30079	HSBCL2A_27006
HSBCL2A_27007	HSBSF2_17080	HSBSF2_17081	HSC4BINDC_40004	HSC5GN_40002	HSCAIII1_26020
HSCEATG_36009	HSCF8N_14077	HSCFOS_11145	HSCN2A_15034	HSCNTFG1_33035	HSCOL301_25041
HSCOLA1I_25034	HSCPB1_25084	HSCPG1_25086	HSCRP01_39004	HSCRPGA_26029	HSCYP450_11121
HSDAFC1_40003	HSDESMIN_33011	HSDHFR01_07056	HSEGFA1_15045	HSEGFA1_15046	HSEGFRG_15043
HSEGFRG_15044	HSENKE_07107	HSEPKER_24002	HSERR_11141	HSFBRGG_11087	HSFCERG5_17084
HSFIBBR1_15029	HSFIXG1_07095	HSFN3_16038	HSG6PD1_30014	HSGASTA_25015	HSGCSFG_17083
HSGG02_11129	HSGLTH1_26027	HSGLUCG2_17067	HSGRFP1_24001	HSH1FNC1_30042	HSH33G1_15024
HSHBB2_11104	HSHISH2A_11068	HSHISH2B_11070	HSHISH3_11073	HSHISH4_11074	HSHL07_07121
HSHMG14A_31007	HSHMG17G_31009	HSHMGCOB_16050	HSHP1G1_11111	HSHSP27_17087	HSHSP70A_17088
HSIFD1_07111	HSIFI54_25038	HSIFI56_25021	HSIFNG_07113	HSIFNIN6_27009	HSIGF2AP_17071
HSIGFIIFE_28010	HSIGK2_07117	HSIL05_07114	HSIL1AG_14063	HSIL1B_17079	HSIL2RG1_11142
HSIL2RG1_11143	HSINSU_07109	HSINV1_16037	HSISGA1_23013	HSKER671_11083	HSLCATG_17058
HSLIPH01_31010	HSLMWOAS_25014	HSLMYC1_27010	HSMDR1A3_35012	HSMHBA1_15054	HSMHCGE1_14076
HSMHCP42_15038	HSMHDC3B_16068	HSMRP14A_26026	HSMRP8A_26025	HSMT1B1_25036	HSMYCC_11146
HSMYCC_11148	HSNEH1PR_40005	HSNFLG_26019	HSNMYC_25008	HSNMYC_25010	HSNRASR_30003
HSOATA_30056	HSOPS_25083	HSP12AA_30062	HSP5301_11223	HSPBGD1_26007	HSPBGD2_26008
HSPEP1_28004	HSPEPC1_28005	HSPGK11_30017	HSPRCA_32001	HSPROL1_14056	HSPS2G1_15056
HSPSBG06_30036	HSRAS1_11149	HSRAS1_16063	HSRAS1_16064	HSRIGA_37014	HSRNBP_39002
HSRPBG1_14047	HSRPS14_24040	HSRPS17A_41007	HSSAACT_25005	HSSISG5B_11139	HSSNRNP3_41008
HSSOD1G1_07053	HSSOMI_16058	HSSP5_25050	HSTCBV81_17093	HSTCR3G1_17096	HSTCRA23_26001
HSTCRT3D_11160	HSTH01_30008	HSTHIO2A_07055	HSTHYR5_14057	HSTKRA_25035	HSTNFA_11158
HSTNFB_11159	HSTNP1_33023	HSTPI_27011	HSTRP_15041	HSTSHBA1_30071	HSTUBAG_14030
HSTUBB2_14031	HSUBILP_28011	HSUG2A_17036	HSURODG_17059	HSVIM5RR_24039	HSVWFB_17050
HSYUBG1_15055	M23631_30077	M24907_29008	M28265_29020	MAAPRTG_25001	MAHMGC01_11116
MAPRP1_26024	MMABLC1A_30026	MMABLC1B_30025	MMACTCA1_29015	MMADAP_15032	MMADIG_17064
MMAGL1_07072	MMALDH1_14050	MMAMY1A1_29011	MMAMY1A2_29013	MMAMY2AI_07097	MMAT01_29019
MMB2ARG_29017	MMBAND31_25042	MMBAND31_25043	MMC31_07093	MMC51_40001	MMCKM1_27013
MMCMDH1_33012	MMCRY1_07069	MMCRYG2D_11085	MMCRYS_11084	MMCSF1PR_37007	MMDH1_32002
MMDHF5_24032	MMENDOA1_16035	MMFABPI_41010	MMFERHG_25047	MMG37_17095	MMGFAPD_14032
MMGLUT1_39003	MMGMCSFG_11138	MMGPD01_24042	MMGSHPX_11120	MMGSTYA1_26004	MMGUSB01_30088
MMH19G_35001	MMHI01_07052	MMHIS2BA_23007	MMHIS2BA_23008	MMHIS2BB_23009	MMHIS2BI_11069
MMHISH31_11071	MMHISH32_11072	MMHPRT1_07058	MMHTF9_17101	MMHTF9_17102	MMIFNBG_23039
MMIG10VH_11151	MMIG19_07118	MMIG31_07120	MMIGHAE_07116	MMIGHAI1_07115	MMIGKAL_07119
MMIGKVH2_29002	MMIGVNP1_14073	MMIL3G_14064	MMIL4G12_15039	MMIL5G_25045	MMKALL_07096
MMLYT22_17097	MMMBP1_27008	MMMDR1_35016	MMMH02_07122	MMMHCC4D_15053	MMMHKBA_14075
MMMOS_29021	MMMOS_29022	MMMP25G1_24028	MMMYBG_15047	MMNPGFI_36011	MMNUCLEO_36015
MMODCAB1_37001	MMOTC1_37005	MMP2AD1_24033	MMPLF42_16059	MMPLP1_30033	MMPOLB_16049
MMPROT2_33029	MMPSP1G_23001	MMRASKI_16065	MMRPL30_11075	MMRPL3A_11076	MMRPOII1_28001
MMRPS16_11077	MMSAA3G1_14039	MMTAT1_14049	MMTHY11G_11161	MMTHYS1_15031	MMTP2A_33024
MMU1A1_17028	MMU1B2_17029	MMU7_36003	MMZFPB1_33039	OABLG1_30028	OAKERC2G_17048
OAKERFG_17047	OCBGLO01_07074	OCCASB5_30006	OCHBAPT_11096	OCUTGLOB_07099	OSCRFA_30037
PTAZGLO_14043	RNAFPG_17052	RNAGPA1H_27014	RNALBA1_28007	RNALDOG1_17061	RNALDOG1_17063
RNAPOA02_30019	RNAPOA4G_30020	RNATROPO_30023	RNCAA_32004	RNCAMI1_23005	RNCASAG1_15030
RNCASGG1_07092	RNCGRPX_29003	RNCPSIA_30055	RNCTRPB_16053	RNCY45E1_07080	RNCYCPRO_27012
RNCYP17G_35052	RNELAI1_29005	RNELAII1_29006	RNFBAG_14035	RNFERL1_25048	RNGLA2U1_17098
RNGROW3_07104	RNHOX_31003	RNHSC73_15051	RNIGF2_25032	RNIGF2_25033	RNIGF2_28008

RNLALB01_07091	RNLHB_30075	RNLPKG_16052	RNLPKG_38002	RNMHCG_16034	RNMLCA1_24037
RNMLCA2_24038	RNMYOLC1_07065	RNOXTNP_07101	RNPBPG_16047	RNPECG1_11117	RNPF4_15040
RNPOLBA_23015	RNPOMCG1_25006	RNPPP_30060	RNPS01_07098	RNPTH2_07108	RNPTRYI_29018
RNRENAA_29023	RNSVFG_14078	RNTHYRP_24029	RNTNTFSG_17046	RNTOG5_11115	RNTRAN_15037
RNU3D_17037	RNVN03_07100	RNWAP1_14040	RRCKBR_30066	RRG33B_30041	RRP450PB_33019
RRRASH_15049	RSANGA1_36005	RSTSHBA1_30069	SGH4H2B_14028	SGHIS1_14026	SGHIS2A3_07048
SGHIS2A3_07049	SGMAPR_41002	SSFSHBS_33036	SSMHCTA_30024	SSPKRIG1_17085	SSUPAG_14053
XBU7SNRNA_36004	XL68KALB_16039	XLACTA2_23037	XLACTCAG_17044	XLAGA1G_11092	XLBGAI_25025
XLBGL3_07078	XLBGLII_25024	XLGFTB_26023	XLGS17A_16093	XLHIS4_07051	XLHISH3G_14027
XLHSP30A_14068	XLHSP70_14069	XLRNU2_17032	XLRPL14_15025	XLRPL1AG_23003	XLTF3A1_14029
XLU5RNA_17042	XLVITE_07088	XLXK81A1_24018			

Appendix C *Drosophila* multiple exon gene data set

Non-redundant multiple exon gene data set ("multi_exon_GB.dat") for *Drosophila melanogaster*, containing 275 gene entries. Constructed as described in the text and public available at http://www.fruitfly.org/sequence/drosophila-datasets.html. Additional background is given on the web site as well.

GenBank Accession	Definition
AB003910	Fruitfly DNA for 88F actin, complete cds.
AF016992	Drosophila melanogaster cytochrome P450 (CYP4D1) gene, complete
AF018964	Drosophila melanogaster B115 cecropin A1 (CecA1) gene, complete
AF018985	Drosophila melanogaster B009 Andropin (Anp) gene, complete cds.
AF018998	Drosophila melanogaster B208 cecropin B (CecB) gene, complete cds.
AF019362	Drosophila melanogaster tailless protein (tll) gene, complete cds.
AF020309	Drosophila melanogaster SNF1A/AMP-activated protein kinase (SNF1A)
AF022650	Drosophila melanogaster kinesin like protein at 38B (Klp38B) gene,
AF025408	Drosophila melanogaster Windbeutel (wind) gene, complete cds.
AF025540	Drosophila melanogaster zinc finger protein HER (her) gene,
AF025792	Drosophila melanogaster 20S proteasome beta2 subunit (beta2_dm)
AF025793	Drosophila melanogaster 20S proteasome alpha7 subunit (alpha7_dm)
AF030334	Drosophila melanogaster smoothened (smo) gene, complete cds.
AF032921	Drosophila melanogaster ribonuclease H1 (rnh1) gene, complete cds.
AF034856	Drosophila melanogaster EF-hand protein NUCB1 (NUCB1) gene,
AF035549	Drosophila melanogaster stress activated MAP kinase kinase 3 gene,
AF035551	Drosophila melanogaster stress activated MAP kinase kinase 4 gene,
AF037336	Drosophila melanogaster antigen 5-related protein (Agr) gene,
AF039233	Drosophila melanogaster MEDEA (Med) gene, complete cds.
AF041048	Drosophila melanogaster CD39-like NTPase gene, complete cds.
AF044925	Drosophila melanogaster hook protein (hk) gene, complete cds.
AF045787	Drosophila melanogaster strain DmL5 fat body protein 2 (Fbp2) gene,
AF047010	Drosophila melanogaster asteroid protein (ast) gene, complete cds.
AF062478	Drosophila melanogaster pyruvate kinase (Pyk) gene, complete cds.
AF068257	Drosophila melanogaster mutL homolog (Mlh1) gene, complete cds.
AF068271	Drosophila melanogaster mutL homolog PMS2 (Pms2) gene, complete
AF069037	Drosophila melanogaster RGS7 gene, complete cds.
AF069297	Drosophila melanogaster pterin-4a-carbinolamine dehydratase gene,
AF069531	Drosophila melanogaster spindle B (spn-B) gene, complete cds.
AF077070	Drosophila melanogaster strain OK13 runt gene, complete cds.
AF079459	Drosophila melanogaster small ras-like GTPase (rab7) gene, complete
AF081252	Drosophila melanogaster mutant fs(2)TW1.RU34 gamma-tubulin
AF086715	Drosophila melanogaster putative histone deacetylase (Rpd3) gene,
CMGCR1A	D.melanogaster GCR 1 gene.
DMACP54	D.melanogaster Acp70A gene, strain M54.
DMAJ4446	Drosophila melanogaster ebony gene.
DMAJ5042	Drosophila melanogaster partial Trl gene and 5'flanking region.
DMALAS	Drosophila melanogaster alas gene.
DMANX	D.melanogaster anxX gene.
DMARIADNE	D.melanogaster ariadne gene.
DMAURG	D.melanogaster aur gene.
DMBAM	D. melanogaster bag-of-marbles (bam) gene, involved in

DMBBBC1	D.melanogaster (Brighton) BBC1 gene.
DMBCDG	Drosophila melanogaster bicoid gene bcd.
DMBJ1G	D.melanogaster BJ1 gene for BJ1 chromatin-binding protein.
DMBOSS	D. melanogaster gene for the bride of sevenless protein.
DMBSG25D	Drosophila bsg25D locus for blastoderm-specific RNA encoding bsg25D
DMBTDGN	D.melanogaster (Canton S) BTD gene.
DMBX200	Drosophila pH200 gene of distal BX-C region (bithorax complex).
DMC49E4	Drosophila melanogaster cosmid 49E4.
DMCALRET	D.melanogaster gene for calreticulin.
DMCHORS16	D. melanogaster gene for chorion protein s16.
DMCOPIAV	Drosophila copia DNA encoding virus-like particle (VLP) protein.
DMCSDUC	D.melanogaster (Canton S) ductin, subunit C proteolipid gene.
DMCYP4D2	D.melanogaster CYP4D2 gene encoding cytochrome P-450.
DMCYSTA	D.melanogaster gene for cystatin-like protein.
DMCZSUDMA	D.melanogaster Cu-Zn superoxide dismutase gene.
DMDEADBXA	D.melanogaster DEAD-box gene, complete CDS.
DMDES1	D.melanogaster des1 gene.
DMDNAJLP	D.melanogaster DnaJ60 gene.
DMDNAMIN	D.melanogaster Minute(2)32A gene.
DMDRCIV2	Drosophila melanogaster DRI class IV gene for type I regulatory
DMDTFIIAS	D.melanogaster dTFIIA-S gene.
DME011778	Drosophila melanogaster beta3 gene.
DME5962	Drosophila melanogaster AP50 gene.
DME9557	Drosophila melanogaster rpS21 gene.
DMEF1AF2	D. melanogaster elongation factor 1-alpha F2 gene.
DMEHAB	D.melanogaster gene for eclosion hormone.
DMELGG	D.melanogaster Elg gene.
DMFBP1	D.melanogaster gene for fat body protein 1.
DMFUSED	D.melanogaster fused gene sequence.
DMGLASS	Drosophila glass gene encoding a zinc finger protein.
DMGTPBP	D.melanogaster gene for GTP-binding protein.
DMH2AVDG	D.melanogaster H2AvD gene for histone H2A variant.
DMH4R	D.melanogaster H4r gene.
DMHAIRG	Drosophila melanogaster DNA for hairy gene.
DMHGSG2	Drosophila heat shock gene 2.
DMJ000880	Drosophila melanogaster gene for mitochondrial porin.
DMK10G	Drosophila K10 gene for putative DNA binding protein.
DMKA12ADH	D.melanogaster (strain KA12) Adh gene for alcohol dehydrogenase.
DMKNIRPS	Drosophila melanogaster zygotic gap gene knirps.
DMKR	Drosophila melanogaster Krueppel gene Kr.
DML2AMD	Drosophila alpha-methyldopa hypersensitive gene l(2)amd.
DMLAMIN	Drosophila gene for lamin.
DMLAMINC	D.melanogaster gene for lamin C.
DMLETHAL2	D.melanogaster gene for male-specific lethal 2.
DMMBNGEN	D.melanogaster gene for lethal(3)malignant blood neoplasm-1 (MBN).
DMMGN	Drosophila melanogaster mago-nashi protein (mgn) gene, complete
DMMP20	Drosophila melanogaster mp20 gene for muscle-specific protein.
DMMSL3	D.melanogaster msl-3 gene.
DMMTNG	Drosophila melanogaster metallothionein gene (Mtn).
DMMTOG	Drosophila Mto gene for metallothionein.
DMNINAA1	Drosophila melanogaster ninaA gene.
DMOHO31	D.melanogaster oho31 gene.
DMORUBCD4	D.melanogaster UbcD4 gene encoding ubiquitin conjugating enzyme.
DMOSBP2	Drosophila melanogaster odorant-binding protein homolog OS-F gene,
DMP11	D.melanogaster gene for P11 protein, A1 related hnRNP.

DMPCGENE	D. melanogaster Pc gene for polycomb protein.
DMPER	Drosophila melanogaster per locus.
DMPGKG	D.melanogaster Pgk gene for phosphoglycerate kinase.
DMPIMP	D.melanogaster pimples gene.
DMPP4	Drosophila melanogaster pp4 gene.
DMPPGENE	D.melanogaster gene for ref(2)Pp protein.
DMPRODOS	Drosophila melanogaster prodos gene.
DMPRUNE	Drosophila melanogaster prune gene.
DMPS35	D.melanogaster PROS-Dm35 gene for 35KDa proteasome subunit.
DMPUFFSP	D.melanogaster mRNA for puff specific protein Bx42.
DMR118C	Drosophila melanogaster intronic R1 gene 18c.
DMRAD54	D.melanogaster RAD54 gene.
DMRAFPO	Drosophila raf proto-oncogene.
DMRLB1A	D.melanogaster Rlb1 gene.
DMRLC1B	D.melanogaster Rlc1 gene.
DMRNPOL2	Drosophila DmRP140 gene for RNA polymerase II 140,000 M(r)subunit.
DMRP128	D.melanogaster DmRP128 gene for RNA polymerase III second-largest
DMRP49	Drosophila gene for ribosomal protein 49 (rp 49).
DMRPA1A	D.melanogaster RPA1 gene.
DMRPL19	D.melanogaster rpL19 gene for ribosomal protein L19.
DMRPL7A	D.melanogaster rpL7a gene.
DMRPS3	D.melanogaster rps3 gene for ribosomal protein S3.
DMSADENO	D.melanogaster gene encoding S-adenosylmethionine decarboxylase.
DMSAL	Drosophila spalt gene, involved in embryogenesis.
DMSGS5	D.melanogaster salivary gland secretion gene Sgs-5 mapping to
DMSPALT	D.melanogaster spalt gene for spalt protein.
DMSPXGENE	D.melanogaster SPX gene.
DMSSRP2GN	D.melanogaster DNA for SSRP2 gene.
DMSTELL	Drosophila stellate gene.
DMSUHW	Drosophila melanogaster su(Hw) gene for suppressor of hairy wing.
DMSWAL	D.melanogaster swallow gene (exons 1, 2 and 3).
DMTFIIB	Drosophila melanogaster Canton S transcription factor IIB (TfIIB)
DMTHR	D.melanogaster thr gene.
DMTID56	D.melanogaster gene encoding Tid(56) protein.
DMTOPII	D.melanogaster gene for type II DNA topoisomerase.
DMTORSO	D. melanogaster torso gene for a putative tyrosine kinase receptor.
DMTOSCAP2	D.melanogaster DNA for Tosca gene.
DMTPIG	D.melanogaster Tpi gene for Triosephosphate isomerase.
DMTRA2W	D.melanogaster tra-2 gene involved in sex determination.
DMTRFG	D.melanogaster TRF gene for TBP-related factor.
DMTROPONI	D.melanogaster tn1 gene.
DMTSLG	D.melanogaster gene for torso-like protein.
DMTU36B	Drosophila TU-36B gene, cytochrome b related protein.
DMU03986	Drosophila melanogaster iso 1 DNA replication inhibitor plutonium
DMU04239	Drosophila melanogaster tolloid (tld) gene, complete cds.
DMU04822	Drosophila melanogaster Oregon R diazepam binding inhibitor (DBI)
DMU06861	Drosophila melanogaster Canton S dihydrofolate reductase gene,
DMU07799	Drosophila melanogaster glutathione-dependent formaldehyde
DMU11718	Drosophila melanogaster TATA-box binding protein (TBP) gene,
DMU15928	Drosophila melanogaster KH-domain putative RNA binding protein
DMU18401	Drosophila melanogaster testis-specific-RRM-protein (Tsr) gene,
DMU19731	Drosophila melanogaster phyllopod (phyl) gene, complete cds.
DMU19742	Drosophila melanogaster vacuolar ATPase subunit A gene, complete
DMU20542	Drosophila melanogaster lethal(1)1Bi protein (l(1)1Bi) gene,
DMU20543	Drosophila melanogaster minute(1)1B protein (M(1)1B) gene, complete

DMU20566	Drosophila melanogaster Ivory Coast isochromosomal line LW8
DMU21218	Drosophila melanogaster zeste-white 4 gene, complete cds.
DMU21552	Drosophila melanogaster CDK5 homolog gene, complete cds.
DMU24676	Drosophila melanogaster twinstar (tsr) gene, complete cds.
DMU27181	Drosophila meiotic recombination and excision repair (mei-9) gene,
DMU34039	Drosophila melanogaster glial cells missing (gcm) gene, complete
DMU35631	Drosophila melanogaster meiotic-218 (mei-218) gene, complete cds.
DMU38951	Drosophila melanogaster vacuolar ATPase subunit E (vha26) gene,
DMU39739	Drosophila melanogaster scarlet protein (st) gene, complete cds.
DMU42204	Drosophila melanogaster putative potassium channel subunit homolog
DMU43588	Drosophila melanogaster geranylgeranyl transferase beta-subunit
DMU43737	Drosophila melanogaster isoaspartyl methyltransferase (Pcmt) gene,
DMU46009	Drosophila melanogaster testes-specific proteasome subunit gene,
DMU49249	Drosophila melanogaster JNK protein kinase (DJNK) gene, complete
DMU49439	Drosophila melanogaster trithorax group protein (ash1) gene,
DMU51043	Drosophila melanogaster alpha esterase (aE1) gene, complete cds.
DMU51045	Drosophila melanogaster alpha esterase (aE3) gene, complete cds.
DMU51046	Drosophila melanogaster alpha esterase (aE4) gene, complete cds.
DMU51047	Drosophila melanogaster alpha esterase (aE5) gene, complete cds.
DMU52952	Drosophila melanogaster CKII beta subunit (CKII-beta1) gene,
DMU56393	Drosophila melanogaster chromosomal protein D1 (D1) gene, complete
DMU59923	Drosophila melanogaster glutamyl-prolyl-tRNA synthetase gene,
DMU60298	Drosophila melanogaster DNA polymerase gamma gene, nuclear gene
DMU63556	Drosophila melanogaster larval serum protein 1 beta subunit
DMU63857	Drosophila melanogaster decapentaplegic protein (dpp) gene,
DMU64721	Drosophila melanogaster 20S proteasome alpha subunit PSMA5 gene,
DMU66357	Drosophila melanogaster ribosomal protein RpL27a gene, complete
DMU66884	Drosophila melanogaster cubitus interruptus dominant protein (ciD)
DMU67905	Drosophila melanogaster rhodopsin 5 (Rh5) gene, complete cds.
DMU69607	Drosophila melanogaster Amyrel gene, complete cds.
DMU78088	Drosophila melanogaster cytochrome P450 (Cyp6a2) gene, complete
DMU83247	Drosophila melanogaster cuticle protein LCP65Ad gene, complete cds.
DMU84745	Drosophila melanogaster cuticle protein LCP65Ac gene, complete cds.
DMU84751	Drosophila melanogaster cuticle protein LCP65Ae gene, complete cds.
DMU84752	Drosophila melanogaster cuticle protein LCP65Af gene, complete cds.
DMU84898	Drosophila melanogaster 14-3-3 epsilon isoform gene, complete cds.
DMUROX	Drosophila melanogaster DNA for urate oxidase (EC 1.7.3.3).
DMW13	D. melanogaster W13 homeobox gene.
DMWHITE	Drosophila melanogaster DNA sequence of white locus.
DMXDH	D. melanogaster Xdh gene for xanthine dehydrogenase (rosy locus).
DMY10276	D.melanogaster stand still gene.
DMYELLOW	Drosophila melanogaster yellow gene.
DMYEMA	D.melanogaster gene for yemanuclein-alpha.
DMYOLK	Drosophila gene for yolk protein I (vitellogenin).
DMYP3G	Drosophila yolk polypeptide gene YP3.
DMZESTE	Drosophila melanogaster zeste gene.
DROAFLL	Drosophila melanogaster GTP-binding protein (arf-like) gene,
DROAPRTZ	Drosophila melanogaster adenine phosphoribosyltransferase (APRT)
DROARF	Drosophila melanogaster GTP-binding protein (ARF-like Arl84F) gene,
DROARF2A	Drosophila melanogaster ADP-ribosylation factor class II (ARF2)
DROARF3B	Drosophila melanogaster ADP ribosylation factor class III (ARF3)
DROARRA	D.melanogaster arrestin (Arr) gene, complete cds.
DROBROWNPR	Drosophila melanogaster brown allele IG281, complete cds.
DROBSHHB	Drosophila melanogaster brain-specific-homeobox protein gene,
DROCDPR	Drosophila melanogaster cdc37 protein gene, complete cds.

DROCOL4G	Drosophila melanogaster collagen type IV gene, complete cds.
DRODCDRK	Fruitfly Dcdrk gene for Dcdrk kinase, complete cds.
DRODEADA	Drosophila melanogaster D-E-A-D box protein (Dbp73D), complete cds.
DRODFUR2X	Drosophila melanogaster Dfurin2 (Dfur2) gene exons 1-16, complete
DRODGQ	D.melanogaster retinal specific G-alpha protein (dgq) gene,
DRODHORO	Drosophila melanogaster dihydroorotate dehydrogenase (dhod) gene,
DRODMRBA	Fruitfly DMR gene for RecA protein homologous, complete cds.
DRODOXA2	Drosophila melanogaster A2 component of diphenol oxidase (Dox-A2)
DRODROSOPH	Drosophila melanogaster serendipity (sry h-l) gene, complete cds.
DRODSOR1	D.melanogaster gene for Dsor1, complete cds.
DROECDINME	Drosophila melanogaster (strain Oregon R) ecdysone-inducible
DROEDG78A	Drosophila melanogaster EDG-78 cuticle protein gene, exons 1 and 2.
DROEDG91A	Drosophila melanogaster EDG-91 gene, complete cds.
DROESCOMBS	Drosophila extra sex combs gene, exon 1-4, complete cds.
DROEST6A	D.melanogaster esterase-6 gene, complete cds.
DROEVE	D.melanogaster even-skipped (eve) gene containing homeo box.
DROFASI	D.melanogaster fasciclin I (FasI) gene, complete cds.
DROGAS02	Drsophila melanogaster G protein alpha subunit gene, complete cds.
DROGLDGMC	D.melanogaster glucose dehydrogenase (GLD) gene, complete cds.
DROGLTFAC	Drosophila melanogaster germline transcription factor gene,
DROHP1	D.melanogaster Hp-1 gene, complete cds.
DROIMPDEH	Drosophila melanogaster inosine monophosphate dehydrogenase gene,
DROLAMAA	Drosophila melanogaster laminin A chain gene, complete cds.
DROLAMB2A	Drosophila melanogaster laminin B2 gene, complete cds.
DROMDR50A	Drosophila melanogaster P-glycoprotein/multidrug resistance protein
DROMEX1A	D.melanogaster mex1 gene, complete cds.
DROMNSO	Drosophila melanogaster manganese superoxide dismutase (mnSOD) gene
DROMSP316	D.melanogaster sex-specific protein msP316 (mst316) gene, complete
DROMYLA	D.melanogaster myosin light chain 2 (MLC-2) gene, complete cds.
DRONANOS	D.melanogaster nanos gene, complete cds.
DRONOD	Drosophila melanogaster kinesin-like protein (nod) gene, complete
DROOPSA	D. melanogaster opsin (ninaE) gene, complete cds.
DROOPSAA	D.melanogaster opsin gene Rh2, complete cds.
DROOSKAR	D.melanogaster oskar gene, complete cds.
DROOTUA	D.melanogaster ovarian tumor protein (otu) gene, complete cds.
DROP40A	D.melanogaster p40 (Stubarista) gene, 5' end.
DROPCNA	D.melanogaster proliferating cell nuclear antigen (PCNA) gene,
DROPCXGEN	Drosophilia melanogaster pcx gene, 5' end.
DROPFK	Drosophila melanogaster phosphofructokinase (pfk) gene, complete
DROPGD	Drosophila melanogaster 6-phosphogluconate dehydrogenase (Pgd)
DROPGLY	Drosophila melanogaster phosphoglycero mutase (Pglym78) gene,
DROPOLA	D.melanogaster POLA gene for DNA polymerase (EC 2.7.7.7) alpha.
DROPOLYABA	Drosophila melanogaster nucleolytic polyadenylate-binding protein
DROPPP	Fruit fly 49-kilodalton phosphoprotein gene, complete cds.
DROPRD	D.melanogaster paired gene (prd) encoding a segmentation protein,
DRORBP1A	Drosophila melanogaster RNA binding protein (rbp1) gene, complete
DRORNAHEL	Drosophila melanogaster DECD family putative RNA helicase gene,
DROROUGH	Drosophila melanogaster developmental protein (rough) gene,
DRORPRIIA	D.melanogaster RpII215 gene encoding RNA polymerase II, largest
DRORPS17	D.melanogaster ribosomal protein S17 gene, complete cds.
DRORPS6X	Drosophila melanogaster ribosomal protein S6 (rps6) gene, complete
DROSEP1HP	Drosophila melanogaster filament protein homolog (sep1) gene,
DROSEV	D.melanogaster sevenless protein gene, complete cds.
DROSNF	Drosophila melanogaster nuclear protein (snf) gene, complete cds.
DROSO7LESA	D.melanogaster son of sevenless gene, complete cds.

DROSSGA	Drosophila melanogaster myosin II (sqh) gene, complete cds.
DROSSL	Drosophila melanogaster casein kinase II beta-subunit homologue
DROSUSG	Drosophila melanogaster suppressor of sable gene, complete cds.
DROTRP	D.melanogaster trp protein gene, complete cds.
DROTUBA1	D.melanogaster alpha-tubulin gene (alpha-1), complete cds.
DROTUBA4	D.melanogaster alpha-tubulin gene (alpha-4), complete cds.
DROVERM	D.melanogaster vermilion protein gene, complete cds.
DROVITB	Drosophila vitelline membrane protein 3C-1 gene, complete cds.
DROXPACDR	Drosophila melanogaster Dxpa gene, complete cds.
DSAJ2740	Drosophila melanogaster capsuleen gene.
U00145	Drosophila melanogaster catalase gene, complete cds.
U00683	Drosophila melanogaster formylglycineamide ribotide
U00790	Drosophila melanogaster proteasome subunit (l(3)73Ai) gene,

Appendix D Drosophila single exon gene data set

Non-redundant single exon gene data set ("multi_exon_GB.dat") for *Drosophila melanogaster* containing 141 gene entries. Constructed as described in the text and public available at http://www.fruitfly.org/sequence/drosophila-datasets.html. Additional background is given on the web site.

GenBank Accession	Definition
AF010325	Drosophila melanogaster CHIP (Chip) gene, complete cds.
AF019020	Drosophila melanogaster B09 diptericin (Dipt) gene, complete cds.
AF030443	Drosophila melanogaster ubiquitin-conjugating enzyme 9 (UBC9) gene,
AF030959	Drosophila melanogaster metchnikowin precursor (Mtk) gene, complete
AF035546	Drosophila melanogaster p38a MAP kinase gene, complete cds.
AF053725	Drosophila melanogaster myristoyl-CoA: protein N-myristoyl
AF069530	Drosophila melanogaster 9 kD basic protein (c550) gene, complete
DM1731L3	Drosophila retrotransposon 1731 3' long terminal repeat (LTR).
DMACTR66B	D.melanogaster AcTr66B gene for actin-related protein.
DMAMYAG1	Drosophila alpha-amylase gene (locus 1) and flanking regions.
DMANGEL	D.melanogaster angel gene.
DMAPTER	D.melanogaster apterous gene for developmental regulatory protein.
DMARM	D.melanogaster arm E16 and E9 genes for armadillo protein.
DMASCT3	Drosophila T3 gene of achaete-scute complex (AS-C).
DMASCT8	Drosophila T8 gene of achaete-scute complex (AS-C).
DMBARI1	D.melanogaster Bari-1 mobile element DNA.
DMBLPP	D.melanogaster (Oregon R) gene for blastopia polyprotein.
DMBX189A	Drosophila distal BX-C region (bithorax complex) pH189 5' region;.
DMC137E7	Drosophila melanogaster cosmid 137E7.
DMCOLT	D.melanogaster colt gene.
DMCSGS	D.melanogaster Cregon R Heidelberg Sgs-4 gene for salivary glue
DMCYCDC3	Drosophila melanogaster cytochrome c gene DC3.
DMCYCDC4	Drosophila melanogaster cytochrome c gene DC4.
DMDC0	Drosophila melanogaster DC0 gene for catalytic subunit of
DMDFD	Drosophila mRNA for Deformed (Dfd) protein.
DMDFR2	D.melanogaster DFR2 gene for FGF-receptor homologue.
DMDISCO	D.melanogaster disconnected (disco) gene for protein causing
DMDM11	D.melanogaster micropia-Dm11 retrotransposon.
DMDOA	D.melanogaster Doa protein kinase gene.
DMDRIBBLE	Drosophlila melanogaster DNA for dribble gene.
DMDROSOCN	D.melanogaster gene encoding antibacterial peptide drosocin.
DME010298	Drosophila melanogaster retrotransposon-like element.
DME010387	Drosophila melanogaster mRNA for beadex/dLMO protein.
DME17355	Drosophila melanogaster PPN 58A gene.
DME223042	Drosophila melanogaster noisette gene.
DMENOLAS	Drosophila gene for enolase (2-phospho-D-glycerate hydrolase) (EC
DMESPLM4	D.melanogaster E(spl) transcription unit m4 gene, enhancer of split
DMESPLM5	D.melanogaster E(spl) transcription unit m5 gene, enhancer of
DMHBG	Drosophila melanogaster hunchback gene encoding a finger protein.
DMHETA1	D.melanogaster Het-A transposable element 17B3 gene for ga-like
DMHISH1	Drosophila melanogaster H1 histone gene.
DMHLHMB	D.melanogaster gene E(Spl)-HLH-mbeta for helix-loop-helix-protein.

DMHOBOG	Drosophila transposable element hobo 108.
DMHS09	D. melanogaster heat shock gene hsp23 with flanking sequences.
DMHSP22G	Drosophila melanogaster gene for heat shock protein hsp22.
DMHSP26G	Drosophila melanogaster gene for heat shock protein hsp26.
DMHSP27G	Drosophila melanogaster gene for heat shock protein hsp27.
DMHSP7	Alpha-gamma fragment from the Drosophila heat shock genes.
DMHSPG3	Drosophila heat shock gene 3 from 67B locus.
DMHSTH331	D.melanogaster mRNA for histone H3.3.
DMIS297	Drosophila melanogaster transposable element 297.
DMKRAKEN	Drosophila melanogaster kraken gene.
DMLYSDG	D.melanogaster LysD gene for lysozyme D.
DMLYSPG	D.melanogaster LysP gene for lysozyme P.
DMMDG3	D.melanogaster mdg3 retrotransposon DNA.
DMMSTBAGE	D.melanogaster gene for protamine (mst35Ba).
DMNESTED	D.melanogaster nested gene for putative Sgs protein.
DMNEU	Drosophila neu mRNA.
DMNG4	D.melanogaster ng-4 gene.
DMNULLOG	D.melanogaster nullo gene.
DMPP1	D.melanogaster gene for protein phosphase 1.
DMRAS3	Drosophila Dras3 gene.
DMREPDNA2	D.melanogaster (Canton S) repeat region DNA.
DMRH92CD	Drosophila R7 photoreceptor cell opsin gene.
DMRPA1G	D.melanogaster rpA1 gene.
DMRPS31	Drosophila melanogaster gene for putative ribosomal protein S31.
DMRUDG	Drosophila melanogaster rudimentary gene.
DMSLP1D	D.melanogaster sloppy paired 1 gene for slp1 protein.
DMSLP2D	D.melanogaster sloppy paired 2 gene for slp2 protein.
DMSPERM	Drosophila Mst87F gene for put. structural sperm protein.
DMSUPFUSE	D.melanogaster suppressor of fused gene.
DMTHB1	Drosophila melanogaster transposon HB1.
DMTPOD	Drosophila DNA for transposable element D near 3'end of dnc gene.
DMTRYAG	Drosophila melanogaster alpha-gene for trypsin-like enzyme.
DMU01335	Drosophila melanogaster ribosomal protein S2 (sop) gene, complete
DMU03277	Drosophila melanogaster Sevelin clone D69 cell cycle arrest protein
DMU03288	Drosophila melanogaster Sevelin clone D52 cell cycle arrest protein
DMU05850	Drosophila melanogaster spatzle (spz) gene, complete cds.
DMU10184	Drosophila melanogaster drosE2F1 protein (drosE2F1) gene, complete
DMU13014	Drosophila melanogaster phosphorylase kinase gamma gene, complete
DMU18130	Drosophila melanogaster masquerade (mas) gene, complete cds.
DMU18307	Drosophila melanogaster heat shock locus (hsr-omega), omega-pre-c,
DMU21123	Drosophila melanogaster ena polypeptide gene, complete cds.
DMU23420	Drosophila melanogaster transposon BEL unknown protein gene,
DMU26939	Drosophila melanogaster arginine kinase (Argk) gene, complete cds.
DMU41064	Drosophila melanogaster putative extracellular ligand trunk gene,
DMU41476	Drosophila melanogaster iota trypsin (iotaTry) gene, complete cds.
DMU42425	Drosophila melanogaster putative Toll-related transmembrane
DMU43583	Drosophila melanogaster kinase suppressor of ras (ksr) gene,
DMU46008	Drosophila melanogaster testes-specific proteasome subunit
DMU52192	Drosophila melanogaster phosphoinositide 3-kinase (cpk) gene,
DMU56257	Drosophila melanogaster intronic protein 259 (IP259) gene, complete
DMU65589	Drosophila melanogaster Dfz2 (Dfz2) gene, complete cds.
DMU71219	Drosophila melanogaster males-absent on the first (mof) gene,
DMU73490	Drosophila melanogaster putative phosphatidyl-inositol-4-phosphate
DMU84749	Drosophila melanogaster cuticle protein LCP65Aa gene, complete cds.
DMU84750	Drosophila melanogaster cuticle protein ACP65A gene, complete cds.

DMU90947	Drosophila melanogaster accessory gland protein Acp76A (Acp76A)
DMU91994	Drosophila melanogaster selenophosphate synthetase (ptuf1) gene,
DMU92867	Drosophila melanogaster bZIP transcription factor (vrille) gene,
DMUBEX	D.melanogaster gene for ubiquitin extension protein.
DMY16065	Drosophila melanogaster a6 gene.
DRO60AP	Drosophila transforming growth factor B-like protein gene (60A),
DROACS1	D.melanogaster achaete gene encoding nerve differentiation,
DROACS2	D.melanogaster scute gene encoding nerve differentiation, complete
DROAMA	D.melanogaster amalgam protein (ama) gene, complete cds.
DROANKY	Drosophila melanogaster ankyrin mRNA, complete cds.
DROCACTUSA	Drosophila melanogaster cactus zygotic protein exons 1-6, complete
DROCLPTN	Drosophila melanogaster calphotin gene, complete cds.
DRODEBB	Drosophila melanogaster membrane-associated protein (deb-b) gene,
DRODEC1A	D.melanogaster defective chorion-1 fc125 (dec-1) gene, complete
DRODSK	D.melanogaster sulfated tyrosine-kinin (DSK) gene, complete cds.
DROESCARGO	Drosophila melanogaster developmental escargot-encoded protein
DROFAT	Drosophila melanogaster fat protein (fat) gene, complete cds.
DROFMRFA2	D.melanogaster FMRFamide neuropeptide gene, exon 2.
DROFREQ	Drosophila melanogaster frequenin gene, complete cds, introns in
DROGADPH1	D.melanogaster glyceraldehyde-3-phosphate dehydrogenase-1 gene.
DROGCL	Drosophila melanogaster (clone 10B-1) germ cell-less protein (gcl1)
DROGGBCS	D.melanogaster (clones T-beta-1E, T-beta-1J) guanine
DROHMGCO	D.melanogaster 3-hydroxy-3-methyglutaryl coenzyme A reductase (HMG
DROHSC4A	Drosophila melanogaster heat shock protein cognate 70 (Hsc4) gene,
DROIMP	Drosophila melanogaster 20-hydroxyecdysone (IMP-E2) mRNA, complete
DROJUN	D.melanogaster Djun gene, complete cds.
DROMSL1A	Drosophila melanogaster male-specific lethal-1 protein (msl-1)
DRONCX	Drosophila melanogaster (clone DR1) Na/Ca exchange protein (NCX)
DRORHO1A	Drosophila melanogaster Rho1 mRNA, complete cds.
DRORNP70K	D.melanogaster U1 70K small nuclear ribonucleoprotein gene,
DRORPIIPD	D.melanogaster RPII215 gene encoding RNA polymerase II subunit, 5'
DROS1C4	D.melanogaster beta-amyloid-like gene, complete cds.
DROSCA	D.melanogaster scabrous protein gene, complete cds.
DROSER1	D.melanogaster serine protease 3 (SER3) gene, complete cds.
DROSIST	sis-Drosophila melanogaster sister-less-a (bZIP) protein gene,
DROSPLEPMC	Drosophila melanogaster split locus enhancer protein mC (E(spl))
DROTFIISAA	Drosophila melanogaster transcription elongation factor (TfIIS)
DROTU4A	D.melanogaster TU-4 mRNA encoding vitelline membrane protein,
DROVITA	Drosophila vitelline membrane protein 26A-1 gene, complete cds.
DROVMP	D.melanogaster vitelline membrane protein gene, complete cds.
S61734S2	omb (biD4)=optomotor-blind gene [Drosophila melanogaster, larvae,
S66940	RPII15=RNA polymerase II subunit 9 [Drosophila melanogaster,
S74038	preprocorazonin (Drosophila melanogaster, Genomic, 680 nt).
SCU41441	Drosophila melanogaster macrolide binding protein (FKBP12) gene,

Chapter 9 Bibliography

Adams, M., *et al.*. (2000). The genome sequence of *Drosophila melanogaster*. *Science* **287**(5461), 2185-95.

Altschul, S. F. & Gish, W. (1996). Local alignment statistics. *Methods Enzymol* **266**, 460-80.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**(3), 403-10.

Andel, F., 3rd, Ladurner, A. G., Inouye, C., Tjian, R. & Nogales, E. (1999). Three-dimensional structure of the human TFIID-IIA-IIB complex. *Science* **286**(5447), 2153-6.

Ashburner, M. & al., e. (1999). European Drosophila Genome Project (EDGP). http://edgp.ebi.ac.uk/.

Ashburner, M., Misra, S., Roote, J., Lewis, S. E., Blazej, R., Davis, T., Doyle, C., Galle, R., George, R., Harris, N., Hartzell, G., Harvey, D., Hong, L., Houston, K., Hoskins, R., Johnson, G., Martin, C., Moshrefi, A., Palazzolo, M., Reese, M. G., Spradling, A., Tsang, G., Wan, K., Whitelaw, K., Kimmel, B. & et al. (1999). An exploration of the sequence of a 2.9-Mb region of the genome of drosophila melanogaster. The *Adh* region. *Genetics* **153**(1), 179-219.

Audic, S. & Claverie, J. M. (1997). Detection of eukaryotic promoters using Markov transition matrices. *Comput Chem* **21**(4), 223-7.

Auger, I. E. & Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bull Math Biol* **51**(1), 39-54.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. (2000). The Pfam Protein Families Database. *Nucleic Acids Res* **28**(1), 263-266.

Baum, L. E. (1972). An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities* **3**, 1-8.

Breathnach, R. & Chambon, P. (1981). Organization and expression of eucaryotic split genes coding for proteins. *Annu Rev Biochem* **50**, 349-83.

Brown, M., Hughey, R., Krogh, A., Mian, I. S., Sjolander, K. & Haussler, D. (1993). Using Dirichlet mixture priors to derive hidden Markov models for protein families. *Ismb* **1**, 47-55.

Brunak, S., Engelbrecht, J. & Knudsen, S. (1991). Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J Mol Biol* **220**(1), 49-65.

Bruskiewich, R., Hubbard, T. & al., e. (1999). GFF-format. http://www.sanger.ac.uk/Software/formats/GFF/.

Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol* **212**(4), 563-78.

Bucher, P. & Trifonov, E. N. (1986). Compilation and analysis of eukaryotic POL II promoter sequences. *Nucleic Acids Res* **14**(24), 10009-26.

Bucher, P. & Trifonov, E. N. (1988). CCAAT box revisited: bidirectionality, location and context. *J Biomol Struct Dyn* **5**(6), 1231-6.

Burge, C. & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**(1), 78-94.

Burge, C. B. & Karlin, S. (1998). Finding the genes in genomic DNA. *Curr Opin Struct Biol* **8**(3), 346-54.

Burley, S. K. & Roeder, R. G. (1996). Biochemistry and structural biology of transcription factor IID (TFIID). *Annu Rev Biochem* **65**, 769-99.

Burset, M. & Guigo, R. (1996). Evaluation of gene structure prediction programs. *Genomics* **34**(3), 353-67.

Cavin Périer, R., Praz, V., Junier, T., Bonnard, C. & Bucher, P. (2000). The Eukaryotic Promoter Database (EPD). *Nucleic Acids Res* **28**(1), 302-303.

Claverie, J. M. (1997). Computational methods for the identification of genes in vertebrate genomic sequences. *Hum Mol Genet* **6**(10), 1735-44.

Claverie, J. M. & Bougueleret, L. (1986). Heuristic informational analysis of sequences. *Nucleic Acids Res* **14**(1), 179-96.

Conaway, R. C. & Conaway, J. W. (1993). General initiation factors for RNA polymerase II. *Annu Rev Biochem* **62**, 161-90.

Demeler, B. & Zhou, G. W. (1991). Neural network optimization for E. coli promoter prediction. *Nucleic Acids Res* **19**(7), 1593-9.

Dunbrack, R. L., Jr., Gerloff, D. L., Bower, M., Chen, X., Lichtarge, O. & Cohen, F. E. (1997). Meeting review: the Second meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CASP2), Asilomar, California, December 13-16, 1996. *Fold Des* **2**(2), R27-42.

Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998). *Biological Sequence Analysis*, Cambridge University Press.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics* **14**(9), 755-63.

Ewing, B. & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**(3), 186-94.

Farber, R., Lapedes, A. & Sirotkin, K. (1992). Determination of eukaryotic protein coding regions using neural networks and information theory. *J Mol Biol* **226**(2), 471-9.

Fickett, J. W. (1982). Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res* **10**(17), 5303-18.

Fickett, J. W. & Hatzigeorgiou, A. G. (1997). Eukaryotic promoter recognition. *Genome Res* **7**(9), 861-78.

Fickett, J. W. & Tung, C. S. (1992). Assessment of protein coding measures. *Nucleic Acids Res* **20**(24), 6441-50.

Fields, C. A. & Soderlund, C. A. (1990). gm: a practical tool for automating DNA sequence analysis. *Comput Appl Biosci* **6**(3), 263-70.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. & Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* **8**(9), 967-74.

Frech, K., Quandt, K. & Werner, T. (1998). Muscle action genes: a first step towards computational classification of tissue specific promoters. *In Silico Biology* **1**.

Friese, E., Reese, M. G. & Rubin, G. M. (1999). *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB), Lyon, France*,

Gelfand, M. S. (1990). Computer prediction of the exon-intron structure of mammalian pre-mRNAs. *Nucleic Acids Res* **18**(19), 5865-9.

Gelfand, M. S. & Roytberg, M. A. (1993). Prediction of the exon-intron structure by a dynamic programming approach. *Biosystems* **30**(1-3), 173-82.

Green, M. R. (1991). Biochemical mechanisms of constitutive and regulated pre-mRNA splicing. *Annu Rev Cell Biol* **7**, 559-99.

Green, P. (1995). *unpublished.*

Grell, E. H., Jacobson, K. B. & Murphy, J. B. (1968). Alterations of genetics material for analysis of alcohol dehydrogenase isozymes of Drosophila melanogaster. *Ann N Y Acad Sci* **151**(1), 441-55.

Guigo, R. (1997). Computational gene identification. *J Mol Med* **75**(6), 389-93.

Guigo, R., Knudsen, S., Drake, N. & Smith, T. (1992). Prediction of gene structure. *J Mol Biol* **226**(1), 141-57.

Harley, C. B. & Reynolds, R. P. (1987). Analysis of E. coli promoter sequences. *Nucleic Acids Res* **15**(5), 2343-61.

Harris, N. L. (1997). Genotator: a workbench for sequence annotation. *Genome Res* **7**(7), 754-62.

Harris, N. L., Helt, G., Misra, S. & Lewis, S. E. (1999). CloneCurator. http://www.fruitfly.org/displays/CloneCurator.html.

Haussler, D. (1998). Computational Genefinding. *Trends in Biochemical Sciences, Supplementary Guide to Bioinformatics*, 12-15.

Hawley, D. K. & McClure, W. R. (1983). Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucleic Acids Res* **11**(8), 2237-55.

Helt, G. & al., e. (1999). Neomorphic Genome Software Development Toolkit (NGSDK). Neomorphic Inc., Berkeley. http://www.neomorphic.com.

Henderson, J., Salzberg, S. & Fasman, K. H. (1997). Finding genes in DNA with a Hidden Markov Model. *J Comput Biol* **4**(2), 127-41.

Hertz, C. B., Krogh, A. & Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Santa Fe Instituted Studies in the Sciences of Complexity. Lecture Notes, Vol 1, 1, Addison Wesley Publishing Company, Santa Fe.

Horton, P. B. & Kanehisa, M. (1992). An assessment of neural network and statistical approaches for prediction of E. coli promoter sites. *Nucleic Acids Res* **20**(16), 4331-8.

Hutchinson, G. B. (1996). The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comput Appl Biosci* **12**(5), 391-8.

Klingenhoff, A., Frech, K., Quandt, K. & Werner, T. (1999). Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* **15**(3), 180-6.

Knudsen, S. (1999). Promoter2.0: for the recognition of PolII promoter sequences. *Bioinformatics* **15**(5), 356-61.

Kohler, R. E. (1994). *Lords of the Fly: Drosophila Geneticsand the Experimental Life*, University of Chicago Press, Chicago.

Kondrakhin, Y. V., Kel, A. E., Kolchanov, N. A., Romashchenko, A. G. & Milanesi, L. (1995). Eukaryotic promoter recognition by binding sites for transcription factors. *Comput Appl Biosci* **11**(5), 477-88.

Kornberg, R. D. (1999). Eukaryotic transcriptional control. *Trends Cell Biol* **9**(12), M46-9.

Krogh, A. (1997). Two methods for improving performance of an HMM and their application for gene finding. *Ismb* **5**, 179-86.

Krogh, A., Ed. (1998). An introduction to hidden Markov models in biological sequences. Computational Biology: Pattern Analysis and Machine Learning Methods. Edited by Salzberg, S., Searls, D. & Kasif, S.: Elsevier.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994a). Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* **235**(5), 1501-31.

Krogh, A., Mian, I. S. & Haussler, D. (1994b). A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Res* **22**(22), 4768-78.

Kulp, D., Haussler, D., Reese, M. G. & Eeckman, F. H. (1996). A generalized hidden Markov model for the recognition of human genes in DNA. *Ismb* **4**, 134-42.

Kulp, D., Haussler, D., Reese, M. G. & Eeckman, F. H. (1997). Integrating database homology in a probabilistic gene structure model. *Pac Symp Biocomput*, 232-44.

Lang, K. J. & Waibel, A. H. (1990). A Time-Delay Neural Network Architecture for isolated Word Recognition. *Neural Networks* **3**, 23-43.

Larsen, N. I., Engelbrecht, J. & Brunak, S. (1995). Analysis of eukaryotic promoter sequences reveals a systematically occurring CT-signal. *Nucleic Acids Res* **23**(7), 1223-30.

Levitt, M. (1997). Competitive assessment of protein fold recognition and alignment accuracy. *Proteins* **Suppl(1)**, 92-104.

Lisowsky, T., Polosa, P. L., Sagliano, A., Roberti, M., Gadaleta, M. N. & Cantatore, P. (1999). Identification of human GC-box-binding zinc finger protein, a new Kruppel-like zinc finger protein, by the yeast one-hybrid screening with a GC-rich target sequence. *FEBS Lett* **453**(3), 369-74.

Lukashin, A. V. & Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26**(4), 1107-15.

Matis, S., Xu, Y., Shah, M. B., Buley, D., Guan, X., Einstein, J. R., Mural, R. J. & Uberbacher, E. C. (1995). Detection of RNA Polymerase II Promoters and Polyadenylation Sites in Human DNA Sequences. *Comput Chem* **20**, 135-40.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* **405**(2), 442-51.

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5**, 115-133.

McLachlan, A. D., Staden, R. & Boswell, D. R. (1984). A method for measuring the non-random bias of a codon usage table. *Nucleic Acids Res* **12**(24), 9567-75.

Michel, C. J. (1986). New statistical approach to discriminate between protein coding and non- coding regions in DNA sequences and its evaluation. *J Theor Biol* **120**(2), 223-36.

Minsky, M. L. & Papert, S. A. (1969). *Perceptrons*, MIT Press, Cambridge.

Moore, M. J., Query, C. C. & Sharp, P. A., Eds. (1993). Splicing of precursors to mRNA by the spliceosome. RNA World, eds. Edited by Gesteland, R. F. & Attkins, J. F. Plainview, NY: Cold Spring Harbor Lab. Press.

Moult, J., Hubbard, T., Bryant, S. H., Fidelis, K. & Pedersen, J. T. (1997). Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins* **Suppl(1)**, 2-6.

Moult, J., Hubbard, T., Fidelis, K. & Pedersen, J. T. (1999). Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins* **Suppl(3)**, 2-6.

Nakata, K., Kanehisa, M. & Maizel, J. V., Jr. (1988). Discriminant analysis of promoter regions in Escherichia coli sequences. *Comput Appl Biosci* **4**(3), 367-71.

Neubauer, G., King, A., Rappsilber, J., Calvio, C., Watson, M., Ajuh, P., Sleeman, J., Lamond, A. & Mann, M. (1998). Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nat Genet* **20**(1), 46-50.

Ohler, U. (1999). Drosophila Promoter Database. .

Ohler, U., Harbeck, S., Niemann, H., Noth, E. & Reese, M. G. (1999). Interpolated Markov chains for eukaryotic promoter recognition. *Bioinformatics* **15**(5), 362-9.

Ohler, U., Stommer, G. & Harbeck, S. (2000). Stochastic Segment Models of Eukaryotic Promoter Regions. *Pac Symp Biocomput* **5**, 377-88.

O'Neill, M. C. (1991). Training back-propagation neural networks to define and detect DNA- binding sites. *Nucleic Acids Res* **19**(2), 313-8.

O'Neill, M. C. (1992). Escherichia coli promoters: neural networks develop distinct descriptions in learning to search for promoters of different spacing classes. *Nucleic Acids Res* **20**(13), 3471-7.

O'Shea-Greenfield, A. & Smale, S. T. (1992). Roles of TATA and initiator elements in determining the start site location and direction of RNA polymerase II transcription. *J Biol Chem* **267**(9), 6450.

Pedersen, A. G. & Engelbrecht, J. (1995). Investigations of Escherichia coli promoter sequences with artificial neural networks: new signals discovered upstream of the transcriptional startpoint. *Ismb* **3**, 292-9.

Penotti, F. E. (1990). Human DNA TATA boxes and transcription initiation sites. A statistical study. *J Mol Biol* **213**(1), 37-52.

Prestridge, D. S. (1995). Predicting Pol II promoter sequences using transcription factor binding sites. *J Mol Biol* **249**(5), 923-32.

Pugh, B. F. (1996). Mechanisms of transcription complex assembly. *Curr Opin Cell Biol* **8**(3), 303-11.

Pugh, B. F. & Tjian, R. (1992). Diverse transcriptional functions of the multisubunit eukaryotic TFIID complex. *J Biol Chem* **267**(2), 679-82.

Qian, N. & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol* **202**(4), 865-84.

Rabiner, L. R. (1989). *IEEE*,

Rabiner, L. R. & Juang, B. H. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine* **3**(1), 4-16.

Reese, M. G. (1994). Erkennung von Promotoren in Pro- und Eukaryontischen DNA-Sequenzen durch kuenstliche neuronale Netze. Diploma Thesis, University of Heidelberg, Germany.

Reese, M. G., Eeckman, F. H., Kulp, D. & Haussler, D. (1997). Improved splice site detection in Genie. *J Comput Biol* **4**(3), 311-23.

Reese, M. G., Harris, N. L., Hartzell, G. & Lewis, S. E. (1999). *The 7th conference on Intelligent Systems in Molecular Biology (ISMB'99), Heidelberg, Germany*, http://www.fruitfly.org/GASP.

Reese, M. G., Hartzell, G., Harris, N. L., Ohler, U. & Lewis, S. E. (2000). Genome Annotation Assessment in *Drosophila melanogaster*. *Genome Res*, **10**(4), 483-501.

Reese, M. G. & Ohler, U. (1999). Gene data sets. http://www.fruitfly.org/sequence/drosophila-datasets.html and http://www.fruitfly.org/sequence/human-datasets.html.

Roeder, R. G. (1996). The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem Sci* **21**(9), 327-35.

Roehrig, S. (2000). Experimental TSS verification for the *unc-86*gene in *C. elegans*. .

Rosenblatt, F. (1962). *Principles of Neurodynamics.*, Spartan, New York.

Rubin, G. M. (2000). Full-length cDNA project. http://www.fruitfly.org/EST

Rubin, G. M. & al., e. (1999). Berkeley Drosophila Genome Project (BDGP). http://www.fruitfly.org.

Rubin, G. M., *et al.*. (2000). Comparative genomics of the eukaryotes. *Science* **287**(5461), 2204-15.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986a). Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing*, Vol. 1 chap. 8.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986b). Learning Representations by Back-Propagation Errors. *Nature* **323**, 533-36.

Sankoff, D. (1992). Efficient optimal decomposition of a sequence into disjoint regions, each matched to some template in an inventory. *Math Biosci* **111**(2), 279-93.

Scherf, M., Klingenhoff, A. & Werner, T. (2000). *in preparation.*

Sharp, P. A. & Burge, C. B. (1997). Classification of introns: U2-type or U12-type. *Cell* **91**(7), 875-9.

Shepherd, J. C. (1981). Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc Natl Acad Sci U S A* **78**(3), 1596-600.

Silverman, B. D. & Linsker, R. (1986). A measure of DNA periodicity. *J Theor Biol* **118**(3), 295-300.

Sippl, M. J., Lackner, P., Domingues, F. S. & Koppensteiner, W. A. (1999). An attempt to analyse progress in fold recognition from CASP1 to CASP3. *Proteins* **Suppl(3)**, 226-30.

Smale, S. T. (1997). Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. *Biochim Biophys Acta* **1351**(1-2), 73-88.

Smale, S. T. & Baltimore, D. (1989). The "initiator" as a transcription control element. *Cell* **57**(1), 103-13.

Snyder, E. E. & Stormo, G. D. (1993). Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks. *Nucleic Acids Res* **21**(3), 607-13.

Snyder, E. E. & Stormo, G. D. (1995). Identification of protein coding regions in genomic DNA. *J Mol Biol* **248**(1), 1-18.

Solovyev, V. & Salamov, A. (1997). The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *Ismb* **5**, 294-302.

Solovyev, V. V., Salamov, A. A. & Lawrence, C. B. (1995). Identification of human gene structure using linear discriminant functions and dynamic programming. *Ismb* **3**, 367-75.

Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* **26**(1), 320-2.

Sonnhammer, E. L., Eddy, S. R. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**(3), 405-20.

Staden, R. & McLachlan, A. D. (1982). Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Res* **10**(1), 141-56.

Stormo, G. D. & Haussler, D. (1994). Optimally parsing a sequence into different classes based on multiple types of evidence. *Ismb* **2**, 369-75.

Thomas, A. & Skolnick, M. H. (1994). A probabilistic model for detecting coding regions in DNA sequences. *IMA J Math Appl Med Biol* **11**(3), 149-60.

Trifonov, E. N. (1987). Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16 S rRNA nucleotide sequences. *J Mol Biol* **194**(4), 643-52.

Trifonov, E. N. & Sussman, J. L. (1980). The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc Natl Acad Sci U S A* **77**(7), 3816-20.

Uberbacher, E. C. & Mural, R. J. (1991). Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc Natl Acad Sci U S A* **88**(24), 11261-5.

Waibel, A. H., Hanazawa, T., Hinton, G. E., Shikano, K. & Lang, K. J. (1989). Phoneme Recognition Using Time-Delay Neural Networks. *IEEE Transactions on Acoustic, Speech, and Signal Processing* **37**(3), 328-39.

Wasserman, W. W. & Fickett, J. W. (1998). Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* **278**(1), 167-81.

Werbos, P. J. (1974). Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. Ph.D. Thesis, Harvard University.

Wiley, S. R., Kraus, R. J. & Mertz, J. E. (1992). Functional binding of the "TATA" box binding component of transcription factor TFIID to the -30 region of TATA-less promoters. *Proc Natl Acad Sci U S A* **89**(13), 5814-8.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pr, M., Reuter, I. & Schacherer, F. (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* **28**(1), 316-319.

Wingender, E., Dietze, P., Karas, H. & Knuppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24**(1), 238-41.

Xu, Y., Einstein, J. R., Mural, R. J., Shah, M. & Uberbacher, E. C. (1994a). An improved system for exon recognition and gene modeling in human DNA sequences. *Ismb* **2**, 376-84.

Xu, Y., Mural, R. J. & Uberbacher, E. C. (1994b). Constructing gene models from accurately predicted exons: an application of dynamic programming. *Comput Appl Biosci* **10**(6), 613-23.

Yokomori, K., Verrijzer, C. P. & Tjian, R. (1998). An interplay between TATA box-binding protein and transcription factors IIE and IIA modulates DNA binding and transcription. *Proc Natl Acad Sci U S A* **95**(12), 6722-7.

Zell, A. & al., e. (1999). Stuttgart Neural Network Simulator (SNNS) 4.2 edit. http://www-ra.informatik.uni-tuebingen.de/SNNS/.

Zemla, A., Venclovas, C., Moult, J. & Fidelis, K. (1999). Processing and analysis of CASP3 protein structure predictions. *Proteins* **Suppl(3)**, 22-9.

Curriculum vitae

Martin G. Reese
3100 College Ave # 4
Berkeley, CA 94705
(510) 597-0567
mgreese@lbl.gov

Personal

Date of Birth	*8 January 1968*
Place of Birth	*Goettingen*
citizenship	*German*

Experience

University of California, Berkeley, CA
Graduate Researcher — *December 1997 - ...*
Berkeley Drosophila Genome Project, Department for Molecular and Cell Biology. Research Advisor: Dr. Gerald M. Rubin

Lawrence Berkeley National Laboratory, Berkeley, CA
Research Scholar — *April 1995 - December 1997*
Human Genome Center, Genome Informatics Group, Research Advisors: Dr. Frank Eeckman and Dr. David Haussler

Neomorphic, Inc., Berkeley, CA
Co-founder — *December 1996*

Technical University of Copenhagen, Copenhagen, Denmark
Research Assistant — *August 1994 - March 1995*
Center for Biological Sequence Analysis and the Physics Department. Research Advisors: Dr. Henrik Bohr, Dr. Jacob Bohr, Dr. Soeren Brunak

Education

Heidelberg University, Heidelberg, Germany
Medical Informatics, Diplom — *1989 - 1994*
Concentration: Mathematical models in medical research (AI-methods)
Grade Point Average: 1.7

Military Service, Germany — *1988*

Kassel Gymnasium, Kassel, Germany
Abitur — *1979 - 1987*
Grade Point Average: 1.3

Selected Publications

Reese, M.G., Harris, N.L., Hartzell, G., Lewis, S.E., (1999). "The challenge of annotating a complete eukaryotic genome: A case study in Drosophila melanogaster", Tutorial at the 7th International Conference on Intelligent Systems for Molecular Biology '99 (ISMB), Heidelberg, Germany, August 6-10, 1999 including the first Genome Annotation Assessment Project (GASP).

Ohler, U., Harbeck, S., Niemann, H., Noeth, E., Reese, M.G., (1999). "Interpolated Markov chains for eukaryotic promoter recognition", Bioinformatics, Vol. 15 Issue 5, 362-369.

Reese, M.G., Eeckman, F.H., Kulp, D., Haussler, D., (1997). "Improved Splice Site Detection in Genie", Journal of Computational Biology, Vol. 4, Issue 3, 311-323.

Kulp, D., Haussler, D., Reese, M.G. and Eeckman, F.H. (1996). "A generalized hidden Markov model for the recognition of human genes in DNA", Proc. Conf. on Intelligent Systems in Molecular Biology '96, St. Louis, Missouri, AAAI/MIT Press.

Reese, M.G., Lund, O., Bohr, J., Bohr, H., Hansen, J.E., and Brunak, S. (1996). "Distance distributions in proteins: A six parameter representation", Protein Engineering, Vol. 9 No. 7.